DYNAMIC SIGNAL OPTIMISATION FOR ISOLATED ROAD JUNCTIONS

by

Woo-Young AHN

A thesis submitted to the University of London
for the degree of Doctor of Philosophy

Centre for Transport Studies
University College London

September 2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The operation of traffic signal has evolved from fixed-time control to traffic-responsive control. The main difference between these two methods is the kind of traffic flow data used for signal timing optimisation. Various techniques have been developed in the literature for traffic-responsive signal control at isolated junctions: non-optimising methods and optimising methods, which differ in their use of detector data and the presence of an objective function. However, the optimising methods developed in previous studies were subject to various simplifications in respect of one or more rules of operation, in ways of interpreting the detector data, specifying the objective function, and making control decisions.

In this study, a systematic approach is presented to develop an optimising method for use in dynamic signal control. Methods of this kind have two main parts: a traffic model and a dynamic optimiser. They are processed in alternating order; the dynamic optimiser provides signal timing plans to the traffic model, and subsequently the traffic model provides estimates of operational performance back to the dynamic optimiser. In this way, the dynamic optimiser makes control decisions, either to extend the current stage or to terminate it. The key feature in this approach is that a plan is developed for the entire lookahead period, but only the first part (one time-step) is implemented. In order to overcome the simplicity of the vertical queueing model, the concepts of kinematics in physics are applied to develop a novel traffic model (the KCS traffic model). Hence, it can be used to interpret detector outputs, and then to estimate three distinct components of delay (detection period delay, prediction period delay and terminal cost) that are used in the objective function. A stage-based exhaustive search method is integrated with the dynamic optimiser within a rolling-horizon formulation. This kind of search method will identify the best plan at that time for the objective in hand, though further arrivals may render that sub-optimal in the future. The lookahead period proposed here is the same for all streams, but varies according to the state of the dynamic optimiser. This optimising method is developed to be used together with various traffic models and objective functions. The formulation is presented on a stream-by-stream basis, so that it is applicable to a wide range of junction configurations.

The performance of this optimising method was compared with existing control methods (System D Vehicle Actuated and fixed-time) by interfacing with the microscopic traffic simulator SIGSIM. The final results were presented in terms of the mean rate of delay with associated standard error to the number of runs. The results show that such a systematic approach can provide substantial improvements in the junction control performance and gives less delay than the existing methods.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Professor Benjamin Heydecker for his constant guidance and encouragement during the whole period of this research. I would also like to thank my secondary supervisor Professor Richard Allsop for his comments and suggestions. Many thanks also go to the examiners of this thesis, Professor M. J. Maher (Napier University) and Professor M. G. H. Bell (Imperial College).

I would like to express my deepest thanks to my wife Eun-Kyung and my parents. It could not be possible to complete this thesis without my father's support.

Finally, many thanks to the members of the Centre for Transport Studies at UCL for their help and warm friendship. Thanks again all the Ph.D. students at Room 222 Chadwich Building, Andy, Dimitris, Ioanna, Patricia, Silvia, Su-Eun, and Taku.

CONTENTS

1. INTRODUCTION 11

1.1 Background 11

1.2 Objectives of the study 13

1.2 Structure of the thesis 15

2. TRAFFIC SIGNAL OPTIMISATION METHODS 17

2.1 Introduction 17

2.2 General Definitions and descriptions 18

2.3 Fixed-time signal control 20

2.3.1 Webster's mathematical frameworks 22

2.3.2 Staged-based optimisation 24

2.3.3 Phase-based optimisation 29

2.4 Traffic-responsive signal control 33

2.4.1 Non-optimising traffic-responsive methods 34

2.4.2 Optimising traffic-responsive methods 35

2.4.2.1 Principles of a dynamic programming approach 35
2.4.2.2 Principles of a rolling-horizon approach 36
2.4.2.3 Optimising methods 38

2.5 Discussion 49

3. TRAFFIC MODELS AND SIMULATION 51

3.1 Introduction 51

3.2 Car-following models 52

3.2.1 General form of car-following models 52

3.2.2 Gipps' car-following model 56

3.2.3 Cellular Automaton (CA) model 58

3.3 Traffic flow models 60

3.3.1 General form of traffic flow models (q - K relations) 60

3.3.2 L-W-R model (Wave model) 61

3.3.2.1 Shock wave model 63

3.3.3 Daganzo's Cell Transmission model 65

3.4 Vertical queueing model 66

3.5 TRANSYT traffic model 68
3.6 SIGSIM simulation model 71
3.6.1 Event-based simulation 72
3.6.2 Vehicle updating and signal control policy 73
3.6.3 Junction geometry and delay 74
3.7 Discussion 75

4. A KINEMATIC CAR-FOLLOWING MODEL FOR SIGNALISED JUNCTIONS 77

4.1 Introduction 77
4.2 Kinematic equations in one-dimension 78
4.3 Trajectory of the leading vehicle 80
4.3.1 Notation for the leading vehicle trajectory and calculation summary 80
4.3.2 The motion up to the braking position ($X_d \rightarrow X_b$) 84
4.3.3 The motion after the braking position ($X_b \rightarrow X_v$) 85
4.3.4 Discussion 88
4.4 Trajectory of the following vehicles 88
4.4.1 The additional notation and calculation summary 88
4.4.2 Departure time at position $\bar{X}_v$ and the motion definition 95
4.4.3 Braking position and time 98
4.4.4 Finding additional variables for the undelayed case 99
4.4.4.1 Crossing time to the stop-line 100
4.4.5 Finding additional variables for the delayed case 100
4.4.5.1 Standard motion test to identify halt case or not-halt case 100
4.4.5.2 Acceleration time, position and speed 102
4.4.5.3 Free-flow position and time 104
4.4.5.4 Crossing time to the stop-line 105
4.4.6 Discussion 107
4.5 Sensitivity of delay at different signal timing plans 108
4.5.1 Delay calculation 108
4.5.1.1 Vertical queueing model delay (vertical delay) 108
4.5.1.2 KCS traffic model delay (KCS delay) 110
4.5.2 Sensitivity analysis 113
4.5.2.1 Sensitivity analysis for the leading vehicle 113
4.5.2.2 Sensitivity analysis for many vehicles 116
4.6 Discussion 119

5. A DYNAMIC OPTIMISATION METHOD 122

5.1 Introduction 122

5.2 Notation 124

5.3 Optimising process 124

5.3.1 Application of exhaustive search method to the optimising process 125

5.4 Delay calculation for the lookahead period of about one-cycle 130

5.4.1 Detection period delay 130

5.4.2 Prediction period delay 133

5.4.2.1 Queue clearance time calculation 134
5.4.2.2 Prediction period delay for the stream currently on green 137
5.4.2.3 Prediction period delay for the stream currently on red 143

5.4.3 Terminal cost function 146

5.4.4 Objective function (Estimates of control performance) 147

5.5 Discussion 148

6. PERFORMANCE ANALYSIS 150

6.1 Introduction 150

6.2 Methodology 150

6.3 Simulation design 151

6.3.1 Input data in SIGSIM 151

6.3.2 Estimation of input parameters for the optimising method 154

6.3.2.1 Saturation flow, start lag, and end lag 155
6.3.2.2 Acceleration rate 157
6.3.2.3 Input parameters in the KCS traffic model and the Vertical queueing model 158

6.3.3 Investigation of vehicle motions 160

6.4 Methods for estimating performance and validation 161

6.4.1 How SIGSIM and the optimising method communicate 162

6.4.2 Two sample means $t-$test and the paired mean $t-$test 163

6.5 Simulation Results 165

6.5.1 Results for balanced inflows 166

6.5.2 Results for unbalanced inflows 172

6.5.3 Sensitivity of control to the maximum green 177

6.6 Discussion 179

7. CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK 181

7.1 Conclusions 181

7.2 Suggestions for further work 184

8. REFERENCES 186

LIST OF FIGURES

Figure 1.1 A diagram for control strategy comparisons in SIGSIM (Optimising method, System D VA and fixed-time) 15

Figure 2.1 Layout of streams 18

Figure 2.2 A typical phasing and staging diagram 19

Figure 2.3 The Rolling-horizon approach 37

Figure 2.4 Principles of estimating the duration of the lookahead period 38

Figure 2.5 Delay changes as used by the MOVA optimiser 47

Figure 3.1 Car-following concepts 53

Figure 3.2 Speed-density curves (Greenberg, Underwood and Eddie) 55

Figure 3.3 Mean speed, wave speed and shockwave speed in $q-K$ curve 64

Figure 3.4 Queue formation and dissipation in the vertical queueing model 68

Figure 4.1 Trajectory of the single vehicle in relation to varying start of green t_g 83

Figure 4.2 Position, time and speed variables for the following vehicle n=2 91

Figure 4.3 Departure time at the downstream position $\overline{X}_v$ 97

Figure 4.4 Standard motion test (halt case and not-halt case motion) 101

Figure 4.5 Trajectory of following vehicles 107

Figure 4.6 Delay for a leading vehicle (Vertical queueing model) 109

Figure 4.7 Delay for a leading vehicle (KCS traffic model) 112

Figure 4.8 Sensitivity of delay for the leading vehicle in varying time t_g 113

Figure 4.9 Sensitivity of delay for the following vehicles in varying time t_g 118

Figure 5.1 Flowchart of the decision making process 123

Figure 5.2 Minimum green and red 125

Figure 5.3 Lookahead period calculation at time t (for the junction has two stages) 126

Figure 5.4 Detection period delay calculation in relation to **t** 131

Figure 5.5 Queue clearance time 135

Figure 6.1 Comparison scenarios 150

Figure 6.2 Junction layout for the simulation 152

Figure 6.3 Saturation flow rate, start lag and end lag estimation in the Road Note 34 method 156

Figure 6.4 Speed reduction against saturation flow in SIGSIM 157

Figure 6.5 Calculation of acceleration rate using the estimated start lag 158

Figure 6.6 Departure time comparison at the stop-line (a fully saturated case: KCS traffic model, Vertical queueing model and SIGSIM) 161

Figure 6.7 SIGSIM and optimising method interface 162

Figure 6.8 A decision tree for choosing the best control strategy according to the paired mean t – test results 166

Figure 6.9 Sensitivity of control to the choice of the maximum green time (non-optimising method VA and optimising method KCS) 178

LIST OF TABLES

Table 4.1 Detector time generation: using shifted exponential distribution of headways 117
Table 6.1 Input parameters in the KCS traffic model and the Vertical queueing model 159
Table 6.2 Departure time of vehicles at the stop-line (fully saturated case) 161
Table 6.3 Mean rate of delay and associated standard error for balanced flows 169
Table 6.4 Two sample means t – test and paired mean t – test for balanced flows 170
Table 6.5 Mean rate of delay and associated standard error for unbalanced flows 174
Table 6.6 Two sample means t – test and paired mean t – test for unbalanced flows 175
Table 6.7 Sensitivity of control to the choice of the maximum green time (mean rate of delay and average green) 178

1. INTRODUCTION

1.1 BACKGROUND

Traffic signals are commonly used to allocate the right-of-way to conflicting streams at urban road junctions, so that the traffic can cross safely and the road system can be used more effectively. Where signals are used, some method is required to calculate durations for their indications. By proper choice of signal timings, the delay to the traffic, the number of stops, and hence the travel time through the system can be reduced. To achieve these aims, a lot of research work has been carried out on the development of isolated junction control and network co-ordinated junction control (or area traffic control). In this study, the focus is put on managing the operation of signals at an individual junction.

The operation of traffic signal has evolved from fixed-time signal control to traffic-responsive signal control. A variety of methods for modern signal timing optimisation are available now, ranging from the simplest two-stage fixed-time control mode to complicated multi-stage traffic-responsive control. The main difference between these two control methods is the kind of traffic flow data used for the calculation of signal timings. In fixed-time signal control, green durations, and hence the cycle lengths are predetermined, often calculated according to a pre-surveyed mean arrival rate, so short-term variations in vehicle arrivals are ignored by the control strategy. However, traffic-responsive signal control uses real-time detector information to extend green durations in responding to varying traffic demands. The aim of this is to take advantage of short-term variations in the arrival patterns of the traffic in order to minimise delays, and hence to improve the quality of traffic flow.

Since the 1930s, most isolated signal junctions in the UK operated with the System D Vehicle Actuated (VA) control (Department of Transport, 1984) until the traffic-responsive MOVA system (Vincent and Young, 1986; Vincent and Peirce, 1988) was developed by the TRRL. The control logic between the approaches of System D VA and MOVA differ in their use of detector data and the presence of an objective function; however, the idea in common to both control methods is to extend the current green until some or all of the detected vehicles have departed. According to the Seoul Metropolitan Police Agency (COSMOS 2001, 2002), 2758 junctions are currently controlled by traffic signals in Seoul, Korea. Among them, 266 are operated by traffic-responsive signal control using the COSMOS system in Seoul, and the rest of them are operated by fixed-time control. Various optimising techniques have been developed in the literature for traffic-responsive signal control and used in many signalised junctions. In general, control decisions are based upon estimates

of current queue length and the time of arrival of traffic which have already been detected (typically 10 seconds: depending on the position of the detector and the desired speed of the link), as well as of traffic that will be arriving during the future timing plans. The VA control uses a gap-seeking method (Van Zuylen, 1976) in which the current green is extended until no vehicles are detected and has no specific objective function for control. Due to a tendency of inefficient green time extensions, and a great sensitivity to inappropriately set maximum green time, this kind of vehicle actuated control is not normally optimal (Bell, Cowell and Heydecker, 1989).

Miller (1963) suggested a self-optimising traffic-responsive control strategy based on the criterion of minimising the total vehicle delay. The lookahead period (planning period) used was untill the end of the next planned green for each stream, and hence it differs between streams. Vehicle arrivals were measured directly from the detectors by assuming a constant travel time from there to the stop-line. Control decisions were made by the current stage extension test, in which the objective function estimated the difference between the delay saved and that lost for a regular integer interval of green extension. No terminal costs were used to represent additional delay incurred beyond the lookahead period. Miller's method has been further studied and modified by many researchers with similar control algorithms, including Bang (1976), Breteque and Jazequel (1979), Gartner (1983), Vincent and Young (1986), Heydecker (1990), and Heydecker and Boardman (1999) amongst others. Gartner's (1983) OPAC used a fixed lookahead period, and hence it was the same for all streams at the junction. Obtaining accurate arrival predictions for this length of time is not feasible in practice with reliability. To reduce this requirement, a rolling-horizon concept was applied to the optimising process. Using this concept, available flow data derived from detectors is used only a part of the horizon period (rollforward period), and the remaining period can be estimated from a suitable model. The key feature in this control method is that an optimal policy is derived for the entire lookahead period, but only the first part of it is implemented, then the lookahead period is rolled by a user-defined time-step. The basic principle in common between Miller and Gartner is seeking the best control decision within a lookahead period. Various optimising techniques have been developed in the literature over forty years, but there are still some features that remain to be investigated in ways of interpreting the detector data, specifying the objective function and making control decisions.

The vertical queueing model that is adopted widely in traffic signal optimisation assumes that all vehicles have the same travel time before joining a queue. Thus, any vehicles crossing a detector have the same travel distance and time to the stop-line, at which vehicles join the queue vertically, not horizontally. The TRANSYT (Robertson, 1969) fixed-time optimisation program has these kinds of constraints in the model, yet it has been shown to give reasonably good estimates of delay in a

road network and to give signal settings which minimise the delay (Whiting, 1972; Kaplan and Powers, 1973; Robertson and Vincent, 1975). The delay is identified in this model as the time difference between a queue departure time and a free-flow arrival time. Once vehicles are held in a vertical queue, the departure time of the first vehicle is assumed to coincide with the start of the effective green time (Clayton, 1940; Webster, 1958; Allsop, 1970). Further following vehicles are assumed to depart the stop-line at equal headways. The vertical queueing model does not consider any braking motion. However, in practice, the motion of vehicles approaching a signal controlled junction is dominated by the current signal displays. The driver will make a decision to either brake, maintain current speed or accelerate according to whether the signal display at the junction is red, green or amber. Consequently, the motion cannot be determined until a signal timing plan is specified.

1.2 OBJECTIVES OF THE STUDY

The main aim of this study is to develop a systematic optimising method for use in dynamic signal control. To accomplish this aim, several smaller objectives have to be specified and integrated with two main parts in the optimising method: a *traffic model* and a *dynamic optimiser*. Against this background, the study objectives are:

1) To overcome the simplicity of the vertical queueing model, a novel traffic model (a Kinematic Car-following model at Signalised junctions: a KCS traffic model) is developed based on the one dimensional kinematic equations of physics. The formulae are derived for two different vehicle groups: a trajectory equation for the leading vehicle and a trajectory equation for following vehicles, in which the motion of vehicles responding to the current signal display is formulated analytically as a function of the start of green time. Hence, in each time-step, the arrival times and departure times of all detected vehicles at the stop-line are estimated using this traffic model. Moreover, the current state of traffic, such as the number of vehicles being detected and the number of vehicles being passed in each stream can be estimated continuously over time. This model is developed to represent individual vehicle motion in relation to the general car-following concepts. Thus, it can be applied in the dynamic signal optimisation at a microscopic level.

2) To make control decisions, a stage-based exhaustive search method is integrated with the dynamic optimiser within a rolling-horizon formulation. That requires use of on-line

detector data and a set of stage sequences. The role of this part is providing the signal timing plans and control decisions to the traffic model. According to an exhaustive search method, all possible combinations of the future signal timing plans are searched between the user-defined minimum and maximum green times. This kind of search method will identify the best plan at that time for the objective in hand, though further arrivals may render that sub-optimal in the future. The lookahead period proposed here is the same for all streams, but varies according to the state of the dynamic optimiser. This optimising method is developed to be used together with various traffic models and objective functions. The formulation is presented on the stream-by-stream basis, so that it is applicable to a wide range of junction configurations.

3) To estimate the operational performance, an objective function is integrated with the traffic model. Here, the traffic model has two choices of trajectory model (the KCS traffic model and the Vertical queueing model) and three choices of objective function ('Detection period delay', 'Detection period delay + Prediction period delay', or 'Detection period delay + Prediction period delay + Terminal Cost'). The role of this part is providing the estimates of delay to the dynamic optimiser. Hence, the dynamic optimiser makes control decisions, whether to 'roll-forward' (i.e. extension of the current stage is beneficial) or 'stage change' (i.e. immediate start of change to the next stage is beneficial). In cases where the roll-forward period is shorter than the detection period, the detector data will be used several times over.

4) To investigate and test the performance of the novel optimising method, the SIGSIM microscopic traffic simulator (Silcock, 1993; Law and Crosta, 1999; Sha'Aban, 2003) is used. Eight different control strategies are tested, including two existing methods: the system D Vehicle Actuated control and fixed-time control (see Figure 1.1). One of the notable feature of SIGSIM is that it includes a measure of 'between runs variability'. SIGSIM has a stochastic element in reliable generation, several runs with the same input data, including identical random number seed, will produce the identical results. However, changing the random number seeds with the same input data will produce different results. This feature is useful, and it is important when comparing different control strategies that a number of runs are carried out to establish the degree of variation arising from random variations in traffic flow patterns. The final results are presented in terms of the mean rate of delay over all streams at the junction with associated standard error to the number of runs. Finally, the estimates of performance for different control strategies are analysed statistically, based on the two sample means $t-$test and the paired mean $t-$test.

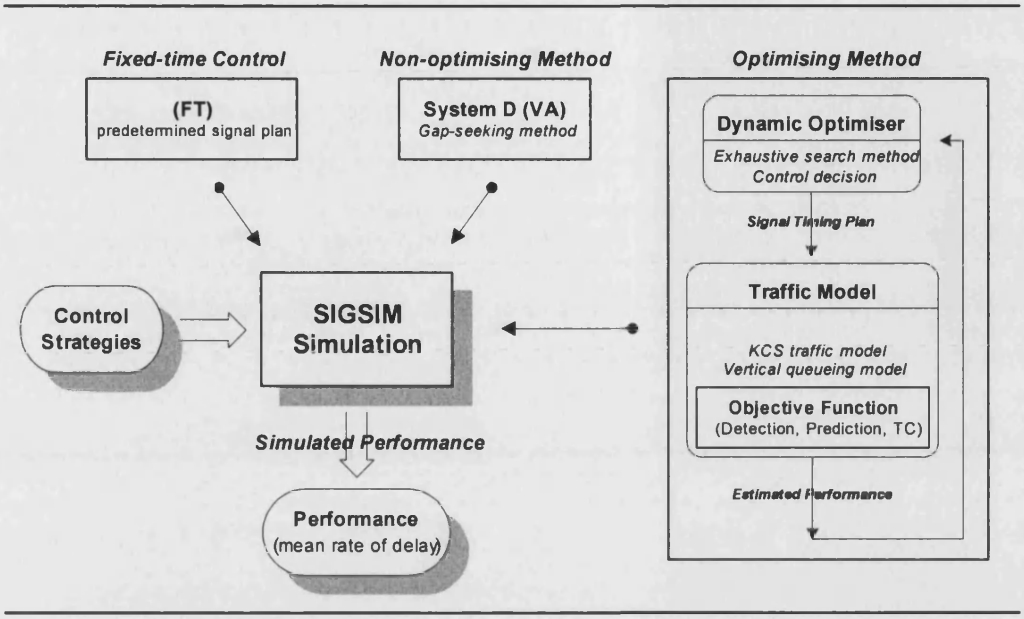


Figure 1.1 A diagram for control strategy comparisons in SIGSIM (Optimising method, System D VA and fixed-time)

1.3 STRUCTURE OF THE THESIS

This thesis has seven chapters. An outline of these chapters is as follows:

Chapter 2 reviews two main types of traffic signal control method for isolated junctions. In fixed-time control, traditional stage-based methods and more flexible phase-based methods are reviewed. In traffic-responsive control, the difference between non-optimising methods and optimising methods are discussed in their use of detector data and the presence of an objective function. Moreover, the advantages and shortcomings between the dynamic programming approach and the rolling-horizon approach are discussed. Some widely known optimising traffic-responsive control methods are presented and discussed in depth.

Chapter 3 presents the necessary knowledge and general background of two main traffic models: car-following models and traffic flow models. Some widely known traffic models are presented. In addition, the characteristics of two computer-based simulation packages, TRANSYT (a mesoscopic model) and SIGSIM (a microscopic model) are discussed in terms of their capabilities for interfacing with the dynamic signal optimisation.

In Chapter 4, a novel traffic model that can be applied to dynamic signal control is proposed. Based on kinematic concepts of physics, a Kinematic Car-following model at Signalised junctions (the KCS traffic model) is developed and formulated analytically as a function of the start of green time. The formulae are classified into two groups: a trajectory for the leading vehicle and a trajectory for a following vehicle. The sensitivity of delay between the simpler vertical queueing model and the more detailed KCS traffic model are compared in relation to varying signal timings.

In Chapter 5, a systematic optimising method for use in dynamic signal optimisation is developed. A stage-based exhaustive search method is integrated with the dynamic optimiser within a rolling-horizon concept is proposed. The respective roles of the traffic model and the dynamic optimiser in the optimising process are discussed. A sequential method for calculating the three distinct components of delay (detection period delay, prediction period delay and the terminal cost) on a stream-by-stream basis is formulated in terms of the light status (currently on green or on red).

In Chapter 6, a performance of the dynamic optimising method is compared with existing methods (System D Vehicle Actuated and fixed-time control) using the SIGSIM microscopic traffic simulator. Methods for estimating representative parameters to be used in the homogeneous numerical equation are presented (e.g. saturation departure time and mean acceleration rate). Eight different control strategies are tested for fourteen different combinations of flow patterns. The simulation results are presented in terms of the mean rate of delay with associated standard error to the number of runs. Finally, the estimates of performance for different control strategies are analysed statistically, based on the two sample means $t-$test and the paired mean $t-$test.

Chapter 7 discusses the main findings of the present study and makes some suggestions for further work.

2. TRAFFIC SIGNAL OPTIMISATION METHODS

2.1 INTRODUCTION

This chapter reviews traffic signal optimisation methods for isolated junctions, the focus in this is a junction at which signal control is not linked with that at adjacent junctions. There are two main types of traffic signal controls that are used in practice: fixed-time signal control and traffic-responsive signal control. Ranging from the simplest two-stage fixed-time control mode to the most complicated multi-stage traffic-responsive control mode, a variety of methods are now available for signal timing optimisation. The main difference between these two control methods is the kind of traffic flow data that are used for optimisation, i.e. whether they are respectively historic flow data or on-line detector data. Here, the historic flow data is the pre-surveyed mean arrival rate, whilst the on-line flow data is the time dependent flow information, which is provided from vehicle detectors.

In fixed-time signal control, green durations and hence the cycle lengths are predetermined according to a pre-surveyed mean arrival rate, in which two or more sets of mutually compatible streams have right-of-way together, but short-term variations in vehicle arrivals are ignored by the control strategy. The operational performance of this method can be estimated conveniently in terms of the pursuing objectives, such as capacity maximisation, cycle time minimisation and delay minimisation (Allsop, 1971, 1972). However, the performance of this approach only can be evaluated after signal timings are specified. In Section 2.3, two main optimisation methods in the fixed-time signal control are discussed: traditional stage-based optimisation methods, and the more flexible phase-based ones.

In traffic-responsive signal control, time dependent detector information is used to take advantage of short-term variations in arrival patterns of the traffic in order to minimise delays and hence to improve the quality of traffic flow. In contrast to the fixed-time signal control, this approach uses the short-term period of detector information to extend green durations in responding to varying traffic demands. The idea behind this approach is to match traffic demands to the durations of the intervals during which streams have right-of-way. On the other hand, the process in such signal control is that a detector provides real-time data, with which a traffic model estimates the current state of traffic, and during which an optimiser decides an optimal timing plan simultaneously. In this way, the performance of the control decision can be evaluated on-line. Various optimising techniques for traffic-responsive signal control have been developed in the literature and currently used in many signal controlled junctions. However, there are still some important features remain to be

investigated in the optimising framework, such as an interpretation of detector data, a specification of the objective function, and a way of making control decisions. In Section 2.4, two different traffic-responsive signal control methods are discussed: non-optimising methods and optimising methods.

2.2 GENERAL DEFINITIONS AND DESCRIPTIONS

The signal sequence in the UK is red, red and amber, green, and amber: the red and amber period starts 2 seconds before the green starts (see Figure 2.2). In contrast, some countries have a sequence of red, green, and amber. In this case, an adjustment of the effective green time definition is necessary. In the literature review, a variety of terminologies can be found in describing and operating a signalised junction, and some terms may not always be used by different authors in the same sense.

As a vehicle approaches a signalised junction, it selects a lane according to the manoeuvre that is interacted. Where one or more adjacent lanes receive identical signal indications and in which traffic queues are formed together, or if vehicles of different movements are sharing one or more adjacent lanes leading to a junction, these vehicles are all considered as a single stream. A ***stream*** is the smallest set of movements that forms a single queue. This is an appropriate segment of traffic to which a queueing analysis is applicable.

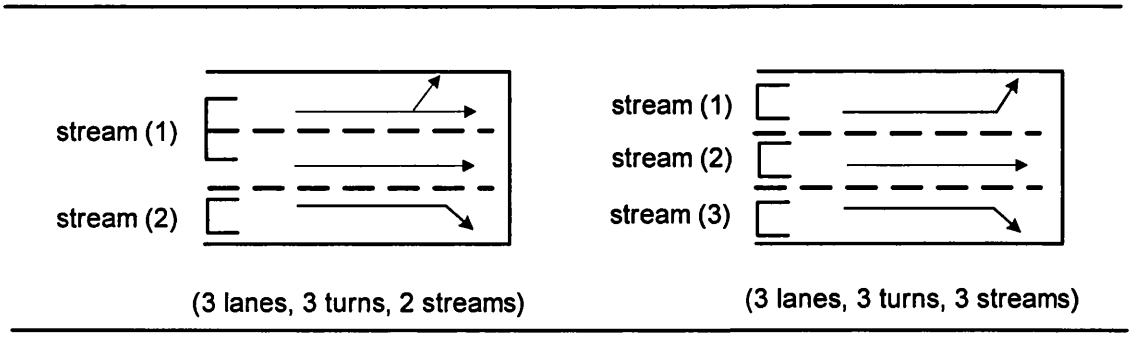


Figure 2.1 Layout of streams

A set of streams that always receives identical signal indications and control together is called a ***group*** or ***phase***. This is the smallest segment of traffic that can be controlled individually. If two groups (phases) can be given green signal indications simultaneously for an indefinite period, then

they are compatible. If two groups (phases) are mutually incompatible, then a clearance time must elapse between phases. The ***clearance time*** between two mutually incompatible groups is the minimum permissible time between the end of the amber of one and the start of the green of the other; this may not be identical for the changes in the two directions. The word phase is sometimes also used to describe the sequence of signal indications received by a set of movements.

The period during which a particular set of mutually compatible phases displays green indications simultaneously while all others display red ones is called a ***stage***. The interval between the end of a green for one phase and the start of green for another conflicting phase is known as the ***intergreen time***. In the UK, this period includes the stopping amber for one phase, which lasts three seconds, and the starting red and amber period for the next phase, which lasts two seconds. The interval period between stages is known as the ***interstage time***, which is based on the intergreen time of the phases.

In fixed-time operation, the change of red, red and amber, green, and amber is periodic. The duration of these is called a ***cycle*** and its duration is the ***cycle time***. A set of signal timings at a junction can be described in two distinct ways, either stage sequence or phase sequence according to whichever is more convenient. In the description of stage sequence, the stage durations are considered as main variables, so the sum of stage durations and interstage lost times consists of one-cycle period.

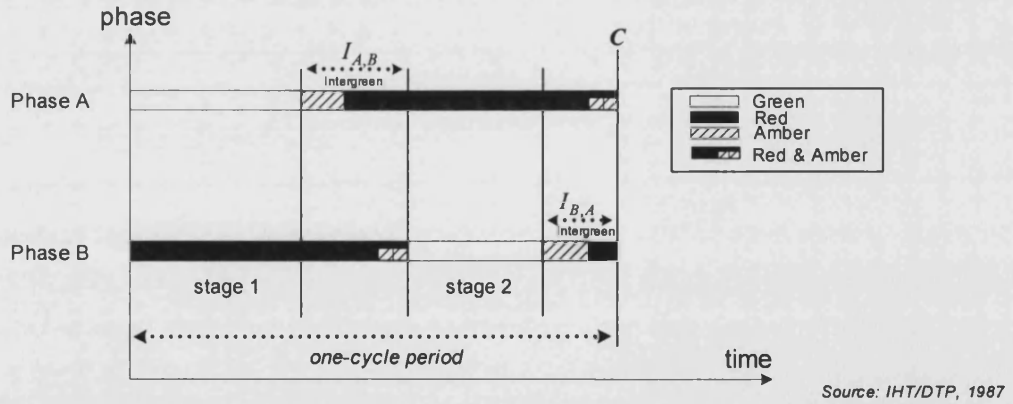


Figure 2.2 A typical phasing and staging diagram

Each stream is associated with it a behaviour of traffic arrivals and departures. The mean number of vehicles arriving into the junction per unit time is known as the ***mean arrival rate***, and vehicles depart from a stationary queue is known as the ***saturation departure rate***. In a signal cycle, time is

usually divided into alternate periods, called effective green time and effective red time. During the ***effective green time***, traffic passes the stop-line at a saturation departure rate while there is a queue, and passes as it arrives if the queue has dissipated. During the ***effective red time***, no vehicles can pass the stop-line. The cycle time is then equal to the sum of the effective green time and the effective red time of each stream.

Capacity at signalised junctions is defined for each stream as the maximum long-term mean rate at which traffic can depart. It is determined as the product of the saturation departure rate and the proportion of effective green time to cycle time. The capacity depends on both the layout of the junction and the signal timings. The ***flow ratio*** for a given lane or stream is defined as the ratio of the actual flow (mean arrival rate) to the saturation flow (saturation departure rate). Then the effective green times are allocated according to the proportion of flow ratio to the cycle time. In addition, the ratio of flow rate to capacity is called the ***degree of saturation***. This is the ratio of the average amount of traffic arriving per cycle to the average amount of traffic passing the stop-line in a cycle throughout which there is a queue.

A common method for assessing the performance of traffic signal control is to estimate the amount of delay incurred by traffic. The delay corresponds to the extra time taken to pass through a junction compared to an unimpeded passage. Two measures of delay are often used at signalised junctions: the mean delay per vehicle and the rate of delay per unit time. The ***mean delay*** over a specific cycle time is given by the total delay divided by the total arrivals in a stream, and has units of seconds. The product of the mean delay and the arrival rate in a stream provides an estimate of the excess number of vehicles in the region of influence due to the operation of signals, which quantity is called the ***rate of delay***, and has units of vehicles (or vehicle hours per hour). In this way, the rate of delay over a specific cycle time is given by the total delay in the cycle divided by the cycle duration. The rate of delay can only be added across all streams at a junction to give the excess number of vehicles in the vicinity of the junction due to the operation of signals, and is known as the ***total rate of delay*** at the junction.

2.3 FIXED-TIME SIGNAL CONTROL

The principle of fixed-time signal control is that the signal timings are predetermined according to the expected flows at the junction. This method applies better to junctions that have a roughly constant arrival flow throughout the time period under consideration, with that flow being within the

capacity. There are two main methods of fixed-time signal control: the stage-based and the phase-based.

One of the first mathematical frameworks for signal timing calculation at a single junction was provided by Webster (1958). The method proceeds by identifying a representative stream for each stage as the one with the greatest flow ratio that runs during that stage. Hence, the green times are allocated into predetermined stage sequences according to the flow ratio of each representative stream to the cycle time. Webster's method is straightforward and gives reasonable signal settings for simple junctions, but are only an approximation to the optimum ones. When the junctions are more complicated, this method is difficult or impossible to use.

A more general approach in signal timing optimisation was developed by Allsop (1971, 1972). He considered all streams at the junction rather than the representative streams, so that this approach is easily applicable to complicated junctions. He formulated the problems of calculating stage durations as a convex mathematical programme to minimise the total rate of delay, and as linear programmes to maximise the reserve capacity and minimise the cycle time. A minimisation problem that involves a convex objective function and a convex feasible region is referred to as a convex optimisation problem. An important feature of the convex optimisation problem is that the optimal solution is unique, so that once a suitable candidate has been identified, it is known to be the only possibility (Luenberger, 1984, p181). If the objective function is linear and the constraints are also linear, then the optimisation problem itself is said to be linear. Methods, such as Webster's and Allsop's use the duration of stages as their main variables that are referred to as *stage-based optimisation*. However, the stage-based formulation has the disadvantage that it requires the user to specify the sequence of stages and the interstage structures before optimisation.

A different approach to cope with the problems of signal setting in stage-based optimisation was proposed by Improta and Cantarella (1984). In this approach, they integrated the problems of optimising phase (group) timings and sequences into a Binary-Mixed-Integer-Linear Program (BMILP), and solved them simultaneously by a branch-and-bound solution method; this allows stage sequences (discrete variables) to be optimised at the same time as the durations of the green indications (continuous variables). However, this method requires that the cycle time be treated exogenously because Webster's (1958) delay function is expressed as a function of the effective green time and the cycle time, in which their formulation is non-linear. This approach has a disadvantage that it only can be applied to linear objective functions.

Heydecker and Dudgon (1987), and Gallivan and Heydecker (1988) developed a related approach from Improta and Cantarella. They directly considered the starting times and durations of green for each group as control variables rather than working through the stages. In cases of more complicated junction layouts, this gives considerable benefits. Most importantly, this method can include the structure of interstage periods and some aspects of cycle time. Thus, the solution can be expressed in terms of the sequence of groups together with their optimised duration and cycle time. Because this method deals directly with groups of streams without the need of maintaining the stage structure during the optimisation, it is called a *phase-based optimisation.* A further discussion of all these approaches has been given by Allsop (1983).They have two important advantages comparing with Improta and Cantarella's method (i.e. linearity and convexity). Improta and Cantarella's control variables have the dimension of time (2.1), whereas Gallivan and Heydecker used proportion of the cycle (2.2). In their control variables, the capacity constraint for a particular stream j is set as follows:

- Improta and Cantarella's capacity constraint is set as

$$g_j \geq \mu c \frac{q_j}{p_j s_j} \tag{2.1}$$

- Gallivan and Heydecker's capacity constraint is set as

$$\Lambda_j \geq \mu \frac{q_j}{p_j s_j} \tag{2.2}$$

where c is the cycle time, q_j is the arrival rate, s_j is the saturation flow rate, $\Lambda_j = g_j / c$ is the proportion of cycle that is effectively green, μ is the common multiplier, and p_j is the maximum acceptable degree of saturation.

In the following sections, Webster's manual method is reviewed first, and two main fixed-time optimisation methods, stage-based and phased-based, are reviewed in terms of the three objectives (capacity maximisation, cycle time minimisation and delay minimisation).

2.3.1 Webster's mathematical frameworks

Methods of calculating signal timings for individual junctions can be divided into two broad classes. In the first class are those methods which can be undertaken manually, whilst in the second are those that rely on computer implementation due to the detailed nature of the calculations involved.

Webster's manual methods (1958) are still used in practice in order to gain familiarity with a junction.

The method proceeds by identifying a representative stream for each stage from those which receive a green indication in it as the one which has the highest flow ratio, which is equivalent to the highest ratio of mean arrival rate to saturation flow rate. Then green times are allocated for each stage as a proportion of representative flow ratio to the available time in the cycle. This method is relatively simple and has been used widely. However, this rule can only be applied when the sequence of stages is sufficiently simple and the associated interstage structures are specified in advance. In particular, the identification of a representative stream in each stage may be difficult or even impossible when some streams have right-of-way in more than one stage.

The optimal cycle time c_0 is calculated from Webster's empirical formula that can minimise the junction delay as a whole is given by

$$c_0 = \frac{1.5L+5}{1-Y} \tag{2.3}$$

where

L is the total lost time in the cycle, calculated as the sum of the durations of the interstage times after allowing for start and end lags, $L = \sum_{i=1}^{m} L_i$,

Y is the sum of the flow ratio of the representative streams in all stages, $Y = \sum_{i=1}^{m} y_i$,

m is the number of stages in the cycle,

y_i is the flow ratio for the representative stream for stage i, $y_i = q_i / s_i$ $(1 \le i \le m)$,

s_i is the saturation departure rate of the representative stream in stage i,

q_i is the mean arrival rate of the representative stream in stage i.

The green durations for stage i are allocated according to the proportion of flow ratio to the effective cycle time. The cycle time (2.3) is split into stages according to the formula

$$g_i = \frac{y_i}{Y}(c_0 - L) \quad (1 \le i \le m) \tag{2.4}$$

where g_i is the effective green time of stage i after start and end lags have been taken into account.

From Webster's manual methods, an important outcome of the capacity analysis is the *volume-to-capacity ratio*, which is also called the *degree of saturation*. The capacity is the maximum rate of vehicles that can be served in a stream at a signalised junction. It depends on the green time available to the traffic and on the saturation flow at the stop-line. The problem of estimating the traffic capacity of a signalised junction is discussed by Webster and Cobbe (1966), and their method is quoted by the Ministry of Transport (1966) along with a nomogram for its use. The allocation of green time (2.4) maximises the reserve capacity of the junction at the given cycle time (2.3). Because the representative streams are the most heavily loaded ones, other streams have lower degrees of saturation. This allocation of green time to stages results in equal degrees of saturation x_i for the representative streams that is given by

$$x_i = \frac{q_i}{Q_i} = \frac{q_i c_0}{s_i g_i} = Y \frac{c_0}{(c_0 - L)} \tag{2.5}$$

where $Q_i = s_i(g_i / c_0)$ is the capacity of the representative stream in stage i, and is determined as the product of the saturation departure rate and proportion of the effective green time to cycle time. This is the maximum throughput in a stage.

Webster found a delay formula by comparison with simulation results, which is reasonably simple in form, and requires measurement of the saturation flow rate and the mean arrival rate alone. The average delay per vehicle passing through the junction is estimated for each stream j. He proposed a two-term delay formula, which includes a factor of 0.9 to correct for the overestimation, that is given by

$$d_j = \frac{9}{10}\left\{\frac{c(1-\Lambda_j)^2}{2(1-\Lambda_j x_j)} + \frac{x_j^2}{2q_j(1-x_j)}\right\} \tag{2.6}$$

where c is the cycle time, $\Lambda_j = g_j / c$, and $x_i = q_j /(\Lambda_j s_j)$ for a particular stream j.

2.3.2 Staged-based optimisation

In stage-based approach, the sequence of stages and the structure of interstages must be specified in advance as an initial step; the cycle time and green duration of each stage are regarded as main control variables. As we mentioned earlier, Webster's methods are limited by the selection of representative streams in a stage; however, Allsop (1971,1972) considers all the streams at the junction rather than the representative streams, so that it is easily applicable to complicated

junctions. He formulated linear programmes to maximise the reserved capacity and minimise the cycle time, and a convex mathematical programme to minimise the junction delay.

Let

m be the number of stages in the cycle,

n be the number of streams at the junction,

a_{ij} be the $m \times n$ stage matrix

$$= \begin{cases} 1 \text{ if stream } j \text{ has right of way in stage } i \\ 0 \text{ otherwise} \end{cases} \qquad (1 \le i \le m,\ 1 \le j \le n)$$

a_{0j} be the proportion of the total lost time that is taken as an extra effective green for stream j, if the stream has green in two consecutive stages this lost time would be used as an extra effective green time, depending on a stage construction $(1 \le j \le n)$,

c be the cycle time,

L be the total lost time per cycle,

λ_0 $= L/c$ be the proportion of the cycle that is taken up by lost time,

λ_i be the proportion of the cycle that is effectively green for stage i $(1 \le i \le m)$,

$g_i^{\min}$ be the minimum green time for stage i $(1 \le i \le m)$,

d_j be the mean delay per vehicle for stream j $(1 \le j \le n)$,

q_j be the mean arrival rate for stream j $(1 \le j \le n)$,

s_j be the saturation departure rate for stream j $(1 \le j \le n)$,

Λ_j be the proportion of the cycle that is effectively green for stream j $(1 \le j \le n)$.

The optimal signal settings can be determined by using mathematical programming. First, we should construct constraints for the sake of safety and efficiency. There are four kinds of constraints that need to be considered in the stage-based optimisation, which are described below:

The signal timings are completely determined by the vector λ. By definition of the lost times and the proportions of a cycle, the *sum constraints* is

$$\sum_{i=0}^{m} \lambda_i = 1 \qquad (2.7)$$

For each stream j, we can identify whether or not the stream j has right-of-way in stage i on the basis of the stage matrix a_{ij} = 1 or 0. Then the effective green time $\Lambda_j c$ is given by

$$\Lambda_j c = a_{0j} L + \sum_{i=1}^{m} a_{ij} \lambda_i c \qquad (1 \le j \le n) \tag{2.8}$$

where $a_{0j} L$ $(0 \le a_{0j} < 1)$ is a known amount of extra effective green time for a particular stream.

Since $\lambda_0 = L/c$, the equation (2.8) becomes

$$\Lambda_j = \sum_{i=0}^{m} a_{ij} \lambda_i \qquad (1 \le j \le n) \tag{2.9}$$

The capacity Q_j of a stream j is given by

$$\begin{aligned} Q_j &= s_j \sum_{i=0}^{m} a_{ij} \lambda_i \qquad (1 \le j \le n) \\ &= s_j \Lambda_j c \end{aligned} \tag{2.10}$$

The effective green time on each stream must be more than great enough to allow all the traffic arriving to pass through the junction in the long run. The flow in a stream should not exceed some portion of capacity given by (2.10), so that $q_j c < s_j \Lambda_j c$. In order to prevent the formation of excessive queues and delays, the *maximum acceptable degree of saturation* p_j is introduced: Webster and Cobbe (1966) suggest p_j = 0.9 in normal circumstances. The *capacity constraints* can be set as

$$\Lambda_j \ge \frac{q_j}{p_j s_j} \qquad (1 \le j \le n) \tag{2.11}$$

If the flow in each stream is multiplied by some common factor μ, then the capacity of the junction as a whole in relation to flows of all streams is estimated. The common value of μ is a measure of the junction capacity for flows proportional to the given arrival rates. The capacity constraints (2.11) can be rewritten as μ in order to find out how much new arrivals can be accommodated. The capacity constraints (2.11) can be rewritten as

$$\Lambda_j \geq \mu \frac{q_j}{p_j s_j} \quad (1 \leq j \leq n) \tag{2.12}$$

If a minimum is imposed on the effective green time for stage i, let it be $\lambda_i c \geq g_i^{\min}$; otherwise, $g_i^{\min} = 0$. By definition of $\lambda_0 = L/c$, the *minimum green time constraints* can be set as

$$\lambda_i \geq \lambda_0 \frac{g_i^{\min}}{L} \qquad (1 \leq i \leq m) \tag{2.13}$$

If a maximum is imposed on the cycle-time, let it be $c_{\max}$, and if a cycle-time is specified, let it be c_0. Then, the *cycle time constraints* can be set as follows:

$$\lambda_0 = L/c_0 \quad \text{(if the cycle time is specified)} \tag{2.14a}$$

$$\lambda_0 \geq L/c_{\max} \quad \text{(if the maximum cycle time is imposed)} \tag{2.14b}$$

where $\lambda_0 \geq 0$.

The optimisation problems can be solved in terms of three intended objective functions: capacity maximisation, cycle time minimisation and delay minimisation. Three such criteria are discussed below:

1) Capacity maximisation

In this case, the quantity of interest is the reserved capacity, in which the capacity will be maximised by allowing the green as long as possible before any of the capacity constraints is violated. If μ>1, then the stream with this signal setting has reserve capacity; otherwise, the stream is overloaded. This is the standard linear programming problem because the objective function and all constraints are linear, thus the solution can obtained by solving the linear programming problem. The objective of this problem is maximising the common multiplier μ subject to the stated constraints:

$$\underset{\mu, \lambda_i (0 \leq i \leq m)}{\text{Maximise}} \ \mu$$

subject to (2.7), (2.12), (2.13) and (2.14).

2) Cycle time minimisation

This is essentially important in the situation of networks of co-ordinated signals where all junctions operate on the same cycle time. A quantity of interest is seeking a shortest cycle time, in which the maximum of $\lambda_0 = L/c$ gives minimum of the cycle time. The objective of this problem is maximising λ_0 subject to the stated constraints. This is also a linear programming problem:

$$\underset{\mu=1,\lambda_i\,(0\le i\le m)}{\text{Maximise}} \quad \lambda_0$$

subject to (2.7), (2.12), (2.13) and (2.14).

3) Delay minimisation

The total mean rate of delay is commonly used to evaluate control performance of a junction. Allsop (1972) has shown that Webster's two-term delay formula is a convex function if it is expressed in the variable of $\lambda_0 = L/c$ and Λ_j. The Webster's 2-term delay formula (2.6) is rewritten as

$$d_j = \frac{9}{10}\left\{\frac{Ls_j(1-\Lambda_j)^2}{2\lambda_0(s_j-q_j)} + \frac{q_j}{2s_j\Lambda_j(\Lambda_j s_j - q_j)}\right\} \qquad (1\le j\le n) \qquad (2.15)$$

By using the equation (2.15), the total rate of delay D for the junction as a whole can be given by

$$D = \sum_{j=1}^{n} q_j d_j \qquad (2.16)$$

A quantity of interest is minimising the total rate of delay for the junction as a whole. Since the objective function is not linear, a nonlinear programming solution method is required. The objective of this problem is minimising D subject to the selected constraints:

$$\underset{\mu=1,\lambda_i\,(0\le i\le m)}{\text{Minimise}} \quad D$$

subject to (2.7), (2.12), (2.13), (2.14), (2.15) and (2.16).

2.3.3 Phase-based optimisation

In the phase-based approach, control variables are used to represent the start times and durations of the green for the phase and cycle, so the explicit stage sequences and interstage structures are not required, but more constraints are required than stage-based. Due to a safety consideration, the intergreen time (or clearance time) for mutually incompatible phases must be specified in advance. In this formulation, the optimisation criteria are the same as for the stage-based optimisation. The problem of applying the phase-based approach to optimal signal timings for a single junction has been discussed by Heydecker and Dudgeon (1987).

Let

n be the number of phases,

θ_j be the proportion of the start of the green for phase j in the cycle time ($1 \le j \le n$),

ϕ_j be the proportion of the duration of the green for phase j in the cycle time ($1 \le j \le n$),

$I_{j,l}$ be the intergreen time period between the mutually incompatible phases, the minimum time required between a phase j loosing a green indication and a phase l gaining one, this value is dependent on the geometric layout of the junction and a particular location of the stop-line.

The cycle time is represented indirectly by using a variable $\xi = 1/c$. By this definition, θ_j and ϕ_j are required to satisfy the following constraints:

$$0 \le \theta_j \qquad (1 \le j \le n) \tag{2.17}$$

$$0 \le \phi_j \le 1 \qquad (1 \le j \le n) \tag{2.18}$$

In (2.17), the value of θ_j will usually between (0,1), but if this value exceeds 1, the fractional part of the resulting value is used. As defined in (2.18), all duration variables are represented in terms of the proportion to the cycle time.

If any minimum and maximum constraints for the green duration and the cycle time are required, these constraints are given as follows:

Let

$g_j^{\min}$ be the minimum acceptable duration of the green for phase j $(1 \le j \le n)$,

$g_j^{\max}$ be the maximum acceptable duration of the green for phase j $(1 \le j \le n)$,

$c_{\min}$ be the minimum acceptable duration of the cycle time,

$c_{\max}$ be the maximum acceptable duration of the cycle time.

The *green time constraints* can be set as

$$g_j^{\min} \xi \le \phi_j \le g_j^{\max} \xi \qquad (1 \le j \le n) \tag{2.19}$$

The *cycle time constraints* can be set as

$$1/c_{\max} \le \xi \le 1/c_{\min} \tag{2.20}$$

The optimisation of these variables is also subject to a number of practical engineering constraints for the sake of safety and efficiency. Hence, every incompatible phases must have an intergreen time (or clearance time) to ensure that the start of phase j is at least one intergreen time later than the end of phase l. Because the optimised performance criteria is depends on orders of incompatible phase groups. As noted, the operation of the signal is periodic, so the choice of the time origin is arbitrary, and any set of mutually compatible phases can receive green first. For any two mutually incompatible phases j and l, let

$$w_{j,l} = \begin{cases} 0 & \text{if the start of green in } (0,c) \text{ for phase } j \text{ prececds that for phase } l \\ 1 & \text{otherwise} \end{cases}$$

Then, $w_{j,l} = 1 - w_{l,j}$ for mutually incompatible phases j and l. The *intergreen time constraints* can be set as

$$\theta_j + \phi_j + I_{j,l}\xi \le w_{j,l} + \theta_l \qquad (1 \le j,\ l \le n) \tag{2.21}$$

From (2.21), if the phase j appears before l, then $w_{j,l} = 0$ and hence $w_{l,j} = 1$ respectively. Thus, the constraint (2.21) can be expressed in two ways:

$$\theta_j + \phi_j + I_{j,l}\xi \le \theta_l \quad \text{and} \quad \theta_l + \phi_l + I_{l,j}\xi \le 1 + \theta_j \tag{2.22}$$

The final constraints are related to the traffic performance of the junction. These ensure that adequate capacity is provided by the signal settings to accommodate the traffic in each stream. This can be set as same concept as the stage-based, but stage sequences are not required. Now we consider the streams k in phase j.

Let

K be the number of streams at the junction,

q_k be the mean arrival rate for stream k $(1 \le k \le K)$,

s_k be the saturation departure rate for stream k $(1 \le k \le K)$,

Λ_k be the proportion of the cycle that is effectively green for stream k $(1 \le k \le K)$,

p_k be the maximum acceptable degree of saturation for stream k $(1 \le k \le K)$.

e_k be the amount of extra effective green time that stream k has in each green interval, including any start and end lags caused by acceleration of vehicles.

The green duration $\Lambda_k c$ for stream k will differ from the phase green duration $\phi_j c$ by an amount of the extra effective green time e_k, so that

$$\Lambda_k = \phi_{g(k)} + e_k\xi \qquad (1 \le k \le K) \tag{2.23}$$

where $g(k)$ is the index of the phase which controls stream k.

The practical capacity constraints are then

$$\Lambda_k \ge \frac{q_k}{p_k s_k} \qquad (1 \le k \le K) \tag{2.24}$$

If the flow in each stream is multiplied by some common factor μ, then the capacity of the junction as a whole in relation to flows of all streams is estimated. The common value of μ is a measure of the junction capacity for flows proportional to the given arrival rates. The capacity constraints (2.24) can be rewritten as

$$\Lambda_k \geq \mu \frac{q_k}{p_k s_k} \qquad (1 \leq k \leq K) \tag{2.25}$$

Once we have set all constraints, the optimisation problems can be solved in terms of three intended objective functions: capacity maximisation, cycle time minimisation and delay minimisation. Three such criteria are discussed below:

1) Capacity maximisation

In this case , the quantity of interest is reserve capacity. Choose $\Phi = (\xi, \theta^T, \phi^T, w^T)$ to maximise μ subject to the stated constraints. The maximum μ can be obtained by solving the linear programming problem:

$$\underset{\Phi}{\text{Maximise}} \quad \mu$$

subject to (2.17)-(2.20), (2.22) and (2.25)

2) Cycle time minimisation

In this case, the quantity of interest is cycle time. Choose $\Phi = (\mu = 1, \theta^T, \phi^T, w^T)$ to maximise $\xi = 1/c$ subject to the stated constraints. This is also a linear programming problem:

$$\underset{\Phi}{\text{Maximise}} \quad \xi$$

subject to (2.17)-(2.20), (2.22) and (2.25)

3) Delay minimisation

Allsop (1972), Murchland (1977) and Gallivan (1982) have shown that Webster's 2-term delay formula is a convex function, if it is expressed in the variables of $\xi = 1/c$ and Λ_k. The Webster's 2-term delay formula (2.6) is rewritten as

$$d_k = \frac{9}{10}\left\{\frac{s_k(1-\Lambda_k)^2}{2(s_k - q_k)\xi} + \frac{q_k}{2s_k\Lambda_k(\Lambda_k s_k - q_k)}\right\} \qquad (1 \leq k \leq K) \tag{2.26}$$

By using the equation (2.26), the total rate of delay D for the junction as a whole can be given by

$$D = \sum_{k=1}^{K} q_k d_k \tag{2.27}$$

In this case, the quantity of interest is delay. Choose $\Phi = (\mu = 1, \theta^{\mathrm{T}}, \phi^{\mathrm{T}}, w^{\mathrm{T}})$ to minimise the total rate of delay. Since the objective function is not linear, a nonlinear programming solution method is required. The objective of this problem is minimising D subject to the selected constraints:

$$\underset{\Phi}{\text{Minimise}} \quad D$$

subject to (2.17)-(2.20), (2.22), (2.25) and (2.27)

2.4 TRAFFIC-RESPONSIVE SIGNAL CONTROL

The principle of traffic-responsive signal control is that the signal timings are not pre-calculated, but are directly influenced by vehicles which are provided by the detectors. Thus, time dependent detector information is used to calculate optimal signal timings. The common aim of this is to take advantage of short-term variations in arrival patterns of the traffic in order to reduce delay and improve the quality of traffic flow. There are two kinds of traffic-responsive signal control method which are used in real junctions:

- Non-optimising traffic-responsive methods.
- Optimising traffic-responsive methods.

Most isolated signal junctions in the UK were operated with the System D Vehicle Actuated (VA) control method (Department of Transport, 1984) until the traffic-responsive MOVA system (Vincent and Young, 1986; Vincent and Peirce, 1988) was developed by the TRRL; the VA is a non-optimising traffic-responsive method, and the MOVA is an optimising traffic-responsive method. The basic principle in common for both control methods is to extend the current stage of green until some or all queues have dissipated, but control logic differs in the utilisation of detector information and the presence of an objective function for control.

The optimisation of signal timings at road junctions in real time is an important and difficult problem, and has received considerable attention in the literature. In the past, various approaches to real time signal optimisation have been proposed by many authors in a variety of ways. Miller

(1963) developed a *self-optimising traffic-responsive strategy* and Robertson and Bretherton (1974) explored backward dynamic programming methods to find the best sequence of signal timings for a known sequence of arrivals. Their methods have been further studied and developed within a dynamic optimisation framework by Bang (1976), Gartner (OPAC, 1983), Henry, Farges and Tuffal (PRODYN, 1983), Vincent and Young (MOVA, 1986), Bell, Cowell and Heydecker (1989), Heydecker (1990), and Heydecker and Boardman (1999) amongst others.

2.4.1 Non-optimising traffic-responsive methods

In general, traffic-responsive signal control methods use information provided from detectors as input data. Optimising formulations use an estimate of delay incurred by the vehicles or some broadly similar quantity as a specific objective function with respect to signal timing plans. However, non-optimising traffic responsive methods do not use an objective function, but decisions are made according to certain rules of operation with data derived from vehicle detectors.

Since the 1930s, most UK isolated signalised junctions have operated with the System D Vehicle Actuated (VA) control of the green times (Department of Transport, 1984); vehicles crossing detectors are the controller (rules of operation) to extend the current green by a certain time beyond a minimum and up to a maximum, or to recall a green if the signal is currently red: this is a gap-seeking method (Van Zuylen, 1976) because a stage will only be terminated when a gap in vehicle detection arises. For the purpose of this method, vehicle detectors are placed at specific points on the approaches to measure vehicle arrivals. These detectors estimate the time when the traffic flow falls below the saturation level. Thus, when the gap between vehicle arrivals at the detector exceeds a certain critical time interval, the current stage will no longer be extended if there are vehicles queueing at conflicting approaches. Otherwise, the current stage will continue to extend by the preset vehicle extension period until its duration reaches the maximum green time.

The VA control has no objective function and runs the controller only using successive vehicle gaps obtained from the detectors. Although such conventional control was often very effective, it does have two limitations:

- A tendency to extend greens inefficiently when traffic flows are less than full saturation rate, especially on multi-lane approaches.
- Great sensitivity of delay to the choice of maximum green times, so that at times of heavy demand this policy extends green to near its maximum duration.

According to Bell, Cowell and Heydecker (1989), this kind of traffic-responsive control is not usually optimal because it does not give the optimal solution for the conditions, especially when a junction configuration is complicated, or maximum green limit is incorrectly set.

2.4.2 Optimising traffic-responsive methods

Optimising traffic-responsive methods are distinguished by their use of a specific objective function within the optimisation framework. The value of the objective function depends on the data that are available from vehicle detectors and the signal timing plans for the near future. The problem of calculating signals at individual junctions in real time has been approached in a variety of different ways. In the *dynamic programming approach*, the control decisions are made by considering possible sequences of signal timings which are generated for a planning period. This has two practical shortcomings: knowledge of arrival patterns during a planning period is needed and the number of possible combinations of decisions to be calculated is in general too large. According to Gartner (1983, 2002), the dynamic programming approach ensures global optimality, but it is unsuitable for on-line use. In the *rolling-horizon approach*, the control decisions are made on the basis of available flow data derived from vehicle detectors. The key feature in this control method is that real-time data is required for only a part of the horizon period, and the remaining period of the horizon can be estimated from a suitable model. If the rollforward period is shorter than the time for which data are available, then the actual arrival data obtained from the detectors are used by the optimiser more than once.

2.4.2.1 Principles of a dynamic programming approach

Dynamic programming (DP) is an optimisation method that can be used for solving certain problems that require a sequence of decisions to be implemented. A decision made at one stage of the problem affects the state of the system afterwards, and hence influences decisions that can be made at subsequent stages. The dynamic programming approach solves this sequential decision problem as a sequence of interrelated decisions. Thus, it can be used for the optimisation of multistage decision processes.

The dynamic programming approach can be applied in signal timing optimisation problems by identifying a suitable state space. This describes each of the traffic states (queue on each approach) and the controller state (current signal settings). The way that the state develops over time depends on arrivals and control decisions, and their interaction, which is estimated using a traffic model.

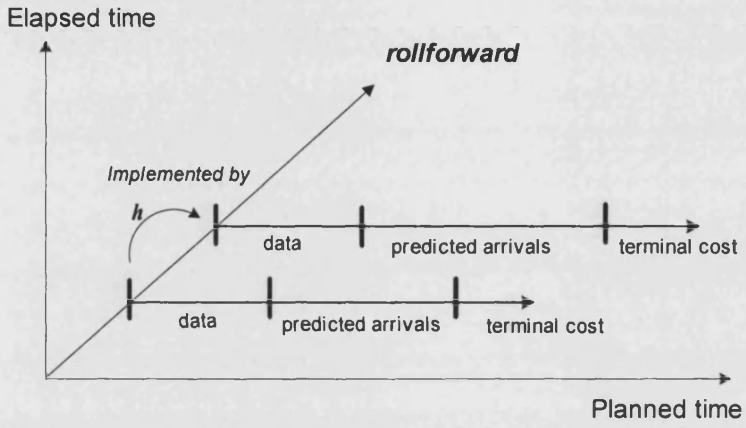
Within the planning period, possible combinations of signal sequences are generated, so that the total delay incurred over a planning period with respect to each sequence of signal timings can be calculated. The set of state that is considered within the optimisation differs according to whether the problem is addressed by backward or forward dynamic programming (BDP or FDP). According to Bell (1989, 1993), the sequence of decisions over the planning period that minimises expected cost is equivalent to finding the shortest path in a certain network; a modified form of Dijkstra's algorithm offers an efficient way to solve the problem. This is a kind of FDP and is more efficient than either forward or backward dynamic programming. The idea of using dynamic programming approaches to optimise signal timings is not new but has previously been considered to be impractical on the two important grounds of data availability and computational burden.

2.4.2.2 Principles of a rolling-horizon approach

In practical signal optimisation, the current state of traffic and queues cannot be measured directly, but can only be estimated using indirect observation from vehicle detectors. As discussed in the previous section, the dynamic programming approach requires complete information of arrivals over the entire control period. To reduce such problems, a rolling-horizon concept was introduced by Gartner (1983). Using this concept, available flow data derived from vehicle detectors are used only a part of the horizon period, and the remaining period can be estimated from a suitable model. Thus, the arrival data obtained from the detectors can be used several times over during the optimising process.

The future interval for which a signal timing plan is calculated and performance evaluated is called the *lookahead period* or *planning period*. Once the optimisation has been undertaken, if the decision is to extend the current stage, then we implement it only for a specific time-step, which is called a *rollforward period*. The key feature in this approach is that an optimal policy is derived for the entire lookahead period, but only the first part of it is implemented then the lookahead period is rolled by a user-defined time-step. The implementation period is usually shorter than the detection period, so arrival data can be used by the optimiser more than once in the optimising process. According to Gartner (2002), the lookahead period is typically taken to be equal to an average cycle length, which is about 1-2 minutes long, and the rollforward period is typically about 2-5 seconds; a shorter roll period implies more frequent calculations and closer to optimum (ideal) results. The relationship between these periods is depicted in Figure 2.3.

In the rolling-horizon approach, the value of the objective function can have three distinct delays: a detection period delay (wherein explicit data are obtained from the detectors), a prediction period delay (wherein arrivals are predicted) and a terminal cost (any residual queues at the end of the lookahead period). More details about these delays will be discussed in Section 5.4.



Source: Bell, 1989

Figure 2.3 The Rolling-horizon approach

The principles for estimating the duration of a lookahead period are based on the queue clearance time of each stream, in which the durations of green are long enough to serve all vehicles in the queue. However, the detailed nature of defining the green durations can differ from method to method. The principles of estimating the duration of lookahead period are shown in Figure 2.4. Consider a junction that has two streams, A and B, in which A is currently on green and B is currently on red. If the current green is extended by h seconds, the duration of the lookahead period can be obtained in respect of the following parameters, where

h	is the rollforward period,
n_A, n_B	is the number of vehicles that will be in the queue for stream A and B respectively, if the current green is extended by h seconds,
q_A, q_B	is the mean arrival rate for stream A and B respectively,
s_A, s_B	is the saturation departure rate for stream A and B respectively,
$l_{A,B}, l_{B,A}$	is the intergreen time from stream A to B, and from stream B to A,

r_A, r_B is the planned red time for stream A and B respectively. It can be obtained by calculation of the average time required to discharge the queue in the opposing stream:

$$r_A = \frac{n_B + s_B l_{A,B}}{s_B - q_B} + l_{B,A} \quad \text{and} \quad r_B = l_{B,A} \frac{n_A + q_A r_A}{s_A - q_A}.$$

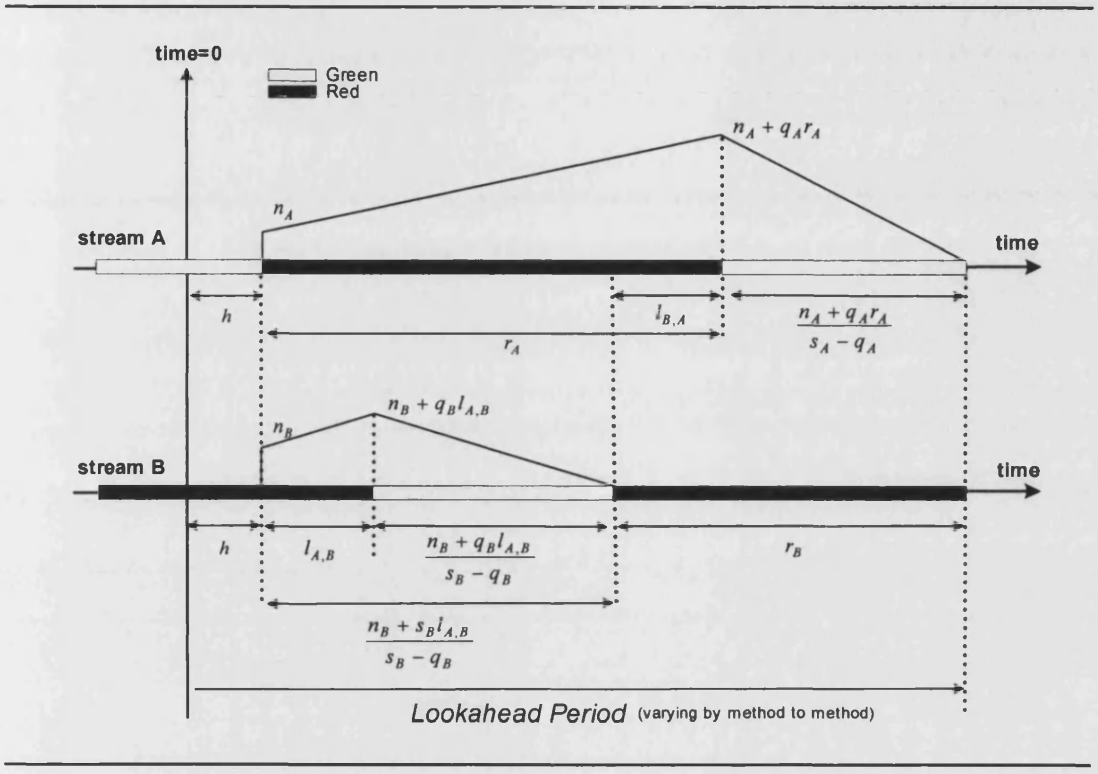


Figure 2.4 Principles of estimating the duration of the lookahead period

2.4.2.3 Optimising methods

Various approaches to the problem of optimising traffic-responsive signal control are reviewed here with reference to dynamic programming and rolling-horizon concepts. In this review, particular attention is paid to self-optimising traffic-responsive signal control methods within rolling-horizon approach:

1) Miller's self-optimising control method

Miller (1963) developed a self-optimising traffic-responsive control method based on the criterion of minimising the total delay of the junction as a whole. The controller scans delays at the junction at

regular intervals of h seconds (e.g. typically every two seconds), and decides to change the signals or to leave them as they are. The decisions are made according to the objective function of minimising the total vehicle delay. This function represents the difference in vehicle-seconds of delay between the saving made by the vehicles that can pass the junction during the h seconds of the current green extension, and the loss to the queuing vehicles in other streams resulting from that extension. At each time of optimisation, the decisions are based on the estimated total delay saving. If the value given by the objective function is positive, then the current green is extended by h seconds; otherwise, the current stage is terminated by calling a new stage.

The lookahead period used by Miller was untill the end of the next green for each stream, and hence it differs between streams. No terminal costs were used to represent additional delays incurred beyond that time due to any residual queues. Vehicle arrivals were measured directly from the detector by assuming a constant travel time from the detector to the stop-line. Arrivals for the prediction period were predicted using an exponentially weighted moving average applied to previous flows.

Miller considered a simple four-stream junction running with two stages, where detectors were placed 100m upstream from the stop-line. He supposed that the first phase for N-S (North and South) was currently on green and the second phase for E-W (East and West) currently on red.

Let

δ_j be the number of vehicles that are expected to pass the stop-line of each stream $j\,(1 \le j \le 4)$; it is possible to anticipate arrivals as far ahead as 10 seconds ($5h$) when the detector is placed at 100 m from the stop-line and approach speed is 10 m/sec,

n_j be the number of vehicles currently waiting on stream j at time of calculation $(1 \le j \le 4)$,

q_j be the arrival rate in h seconds for stream $j\,(1 \le j \le 4)$,

s_j be the saturation departure rate in h seconds for stream $j\,(1 \le j \le 4)$,

r_i be the duration of planned red time for stage i $(1 \le i \le 2)$, the prediction formula of the next red time that Miller supposed was $r_{NS} = 10 + (n_E + n_W)$,

a be the amber time,

l_i be the lost time for stage i $(1 \le i \le 2)$.

In his performance formula, Miller considered the delay in two parts: the first one for the delay saving due to N-S green extension by h seconds, and the second one for the delay caused by E-W red extension by h seconds. Miller used the number of affected vehicles in his delay formula, which are estimated in relation to the current queue dissipation time and the saturation departure rate.

By extending the current N-S green by h seconds, the start of the next N-S green is put back by at least h seconds, but it is not always true when traffic flows are light. If the N-S green is extended by h seconds, δ_j number of vehicles can pass through the junction, but as well as q_j number of new arrivals stay in the queue for an extra h seconds until the signals change in their favour again. In order to calculate the delay saving for the current green extension, Miller used number of *affected vehicles*, which can be obtained by multiplying the queue dissipation time and the saturation departure rate. If δ_j number of vehicles is in the queue, that affects $(\delta_j s_j / s_j - q_j)$ number of departures whilst δ_j are dissipating; in this way, q_j number of vehicles in the queue affects $(q_j s_j / s_j - q_j)$ number of departures. Here, the factor $s/(s-q)$ takes into account the continuous build-up of the new queue during the existing queue discharge (Bang, 1976). The saving of delay to the N-S traffic by extending the green by h seconds is given by

$$\left[\delta_N - q_N\left(\frac{1-\delta_N/s_N}{1-q_N/s_N}\right)+\delta_S - q_S\left(\frac{1-\delta_S/s_S}{1-q_S/s_S}\right)\right](a+r_{NS}+l_{NS}) \qquad (2.28)$$

The Miller's original equation (2.28) can be rewritten as

$$\left[\frac{s_N(\delta_N-q_N)}{s_N-q_N}+\frac{s_S(\delta_S-q_S)}{s_S-q_S}\right](a+r_{NS}+l_{NS}) \qquad (2.29)$$

The second part of the performance formula looks for the delay caused by the red extension. In this case, Miller used new arrivals during the lookahead period. As seen in Figure 2.4, when the signals change in favour of the E-W traffic (that is stream A in Figure 2.4), after the lost time has elapsed, the formed queue during the current red will start to dissipate at the time when the green comes on. Each stream may have the different queue clearance time. Thus, in Miller's method, the lookahead period is different for each stream.

At the time of the optimisation, Miller predicted the growth of the queue during the lost time, i.e. the initial queue plus arrivals during the lost time, which would take $(n_j + q_j l_i)/(s_j - q_j)$ time to clear. For a stage E-W, the total number of affected vehicles for an extra h seconds is given by

$$[(n_E + q_E k_E) + (n_W + q_W k_W)] \tag{2.30}$$

where $k_E = l_{EW} + \frac{n_E + q_E l_{EW}}{s_E - q_E} = \frac{n_E + s_E l_{EW}}{s_E - q_E}$ is the lost time plus queue clearance time for the stream E, and respectively $k_W = \frac{n_W + s_W l_{EW}}{s_W - q_W}$ for the stream W.

Using the equation (2.29) and (2.30), Miller's test quantity T is given by

$$T = \left[\frac{s_N(\delta_N - q_N)}{s_N - q_N} + \frac{s_S(\delta_S - q_S)}{s_S - q_S}\right](a + r_{NS} + l_{NS}) - \left[(n_E + q_E k_E) + (n_W + q_W k_W)\right]h \tag{2.31}$$

At each time of optimisation, the test quantity T gives the difference between the delay saved by the green extension for the approach N-S, and the additional delay caused by that for the approach E-W. If this quantity is positive ($T>0$), this means that it would be advantageous to extend the current stage; otherwise, it indicates that no delay saving will be achieved by extending the current stage. Miller had noticed that in some such cases delay saving could be achieved by extending green for longer times. He therefore suggested that extension tests should be done for the next eight or ten seconds ($4h$ or $5h$).

In a review, Bell, Cowell and Heydecker (1989) commented that Miller's method takes advantage of any short-term variations in flow in order to reduce delay. However, as Miller noted that for the periods of heavy demand, this method would result in the allocation of long green times to the stream with greater saturation departure rate and this might not be acceptable in practice. According to Breteque and Jezequel's (1979) field tests, Miller's algorithm showed an average saving of travel time of about 10 seconds (15% per vehicle for moderate traffic) in comparison with the optimised fixed-time.

2) Robertson and Bretherton's DYPIC

Robertson and Bretherton (1974) explored the problem of optimising the sequence of signal setting at a simple two-stream junction with known arrival patterns. They used a backward dynamic programming (BDP) approach to identify the best sequence of control decisions at 5 seconds resolution, which provides minimum delay over a 600 second planning period through which the arrival pattern of vehicles is arbitrary but predetermined. Each state has a arrays of minimum delay, which are dimensioned with respect to the maximum number of vehicles that can be held in a queue. These state arrays are updated based on the signal decisions and arrivals by working backwards from the end of planning time T to time t=0. The significance of this method is that if the minimum delay and optimal policy for the remainder of the planning period are known for all queue and signal states at time t, this information can be used to calculate the minimum delay and optimal policy at time (t-1). The procedure is completed at time t=0 with the calculation of minimum delay arrays and the optimal policy.

On this basis, Robertson and Bretherton showed that the total delay incurred at a simple cross-road under optimal control during a 600 seconds lookahead period could be approximated closely by an expression that is quadratic in initial queue sizes. This value was then added at the end of the detection period as a terminal costs (see Section 5.5.3). The principle of BDP was used to find the optimal control policy and the method was implemented by a computer program called DYPIC (Dynamic Programmed Intersection Control). They recognized that this method was impractical for implementation as a traffic-responsive method due to data availability and computational burden, but this does show that use of a 10 second detection period and terminal cost functions can result in good performance.

3) Bang's control method after Miller's

Bang (1976) developed a similar self-optimising control strategy to Miller's. He named this TOL (Traffic Optimisation Logic), and compared it to the vehicle actuated system D (VA) control using field and simulation tests. Field tests were performed at one urban and one suburban junction in Stockholm. The result of the field tests indicated that the TOL strategy gave 20-25% reduction of average delay as compared with VA. These finding are consistent with field observations report by Breteque and Jezequel (1979).

As in Miller's strategy, the green extension in TOL is based on the calculations of delay at regular intervals of h seconds (h=1 seconds) represented by an objective function. This function represents

the saving or loss resulting from extension of the prevailing green time by h seconds. Like Miller, Bang considered two parts to the delay, but they differed in the number of vehicles that they considered in the formula: the first part used the number of vehicles that could cross the stop-line if the green were extended by h seconds, and the second part used the number of vehicles affected in relation to the current queue clearance time.

The method is exemplified for the simple case of the junction between two one-lane approaches A and B. Suppose that stream A is currently green and stream B is currently red, and hence the decision to be made is whether or not to extend the current green A, based on the evaluation of the control function T_A.

Let

$n_A(t), n_B(t)$	be the number of vehicles in stream A and stream B respectively at the time of optimisation,
$\delta_A(t)$	be the number of vehicles that can pass the junction if the green is extended by h seconds for stream A at the time of optimisation,
q_A, q_B	be the mean arrival rate for stream A and B respectively,
s_A, s_B	be the saturation departure rate for stream A and B respectively,
$I_{A,B}, I_{B,A}$	be the intergreen time from stream A to B, and one from stream B to A respectively,
r_A	be the planned red time for stream A, which can be obtained by calculation of the time required to discharge the queue in the conflicting stream B.

At the time of optimisation, vehicle arrivals are directly measured by using *long-loop detectors*, in which the arrivals are estimated based upon the average speed v_0, the length of the long-loop detector LL and number of vehicles $n_A(t)$ occupied within the long-loop detector. The predicted number of vehicles $\delta_A(t)$ that can pass the stop-line in the green extension by h seconds is given approximately by

$$\delta_A(t) = \frac{n_A(t) h v_0}{LL_A} \qquad (t \text{ is the time of optimisation}) \qquad (2.32)$$

As seen in Figure 2.4, at time t=0, if the stream A green is extended by h seconds, a growth of queue in stream B is estimated after overlapping the intergreen time period $I_{A,B}$. The growth of queue in stream B during the intergreen time can be predicted based on the upstream flow q_B. The growth of queue $\tilde{n}_B$ in the stream B due to the h seconds extension is estimated by

$$\tilde{n}_B = n_B(t) + q_B I_{A,B} \qquad (t \text{ is the time of optimisation}) \tag{2.33}$$

The affected number of vehicles to be discharged whilst the queue formed in (2.32) is dissipating can be estimated by

$$n_B(t) = \frac{\tilde{n}_B s_B}{s_B - q_B} \qquad (t \text{ is the time of optimisation}) \tag{2.34}$$

The duration of the forthcoming red time for stream A can be obtained by the time required to discharge the queue in stream B. This is equal to the total intergreen time plus the time taken for $\tilde{n}_B$ vehicles to dissipate. The duration of the next red time r_A for the stream A is thus estimated by

$$r_A = I + \frac{\tilde{n}_B}{s_B - q_B} \tag{2.35}$$

where I is the sum of the intergreen times, $I = I_{A,B} + I_{B,A}$.

The decision of whether or not to extend the current stage in the TOL strategy is based on calculations of T_A or T_B at regular intervals h. The control function T_A at time t is given by

$$\begin{aligned} T_A(t) &= r_A \delta_A(t) - h n_B(t) \\ &= \left[(I + \frac{n_B(t) + I_{A,B} q_B}{s_B - q_B})(\frac{n_A(t) h v_0}{L L_A}) \right] - h \left[\frac{s_B (n_B(t) + I_{A,B} q_B)}{s_B - q_B} \right] \end{aligned} \tag{2.36}$$

4) Gartner's OPAC

Gartner (1983) developed the OPAC (Optimisation Policy for Adaptive Control) system which includes three different control strategies: a dynamic programming approach, a simplified

(sequential optimisation) approach and a rolling-horizon approach. Unlike DYPIC, this was intended for real-time signal control of networks.

The initial version of OPAC uses the BDP approach, so that it produces an optimal sequence of timings for any given horizon if accurate information is available on arrivals for the entire control period. However, this method cannot be used for real-time implementation because of the extensive computational effort as well as the difficulty of obtaining sufficient real-time information in practice. Therefore, a simplified approach was developed which achieves most of the performance of BDP, and overcomes its extensive computational burden, so that it becomes suitable for on-line implementation. Gartner limited the number of stage changes (at least one signal change and up to three) that would be considered during the lookahead period. The optimal stage change policies are calculated for each stage in a forward sequential manner for the entire lookahead period. However, this approach also requires knowledge of arrivals over the entire planning period, which is difficult to obtain in practice with reliability. To reduce these requirements, he introduced a rolling-horizon approach to make use of data as they became available from detectors: he showed that this can be achieved without substantial degradation in performance.

The method was applied to a simple junction similar to that of Robertson and Bretherton, using a 5 seconds resolution time, upstream detectors that provide advance flow information for the detection period of 15 seconds (rollforward 15 seconds), constant arrivals for the remainder of the 60 seconds lookahead period, and planning up to three changes of stage. This form of OPAC was found to perform 30-50 per cent better than fixed-time control. Further investigations were undertaken by Chen, Cohen, Gartner and Liu (1987), in which the OPAC method was implemented and tested using the NETSIM (FHWA, 1982) simulation tool. With more realistic junction configurations, it showed 5 per cent better performance than vehicle-actuated (VA) control and 10 per cent better than optimised fixed-time control.

5) Henry, Farges and Tuffal's PRODYN

Henry, Farges and Tuffal (1983) adopted another approach to developing a practical control strategy from a dynamic programming approach similar to those of DYPIC and OPAC. The PRODYN method uses a forward dynamic programming (FDP) algorithm in place of BDP. Like OPAC, PRODYN was designed for real-time signal control of networks.

The control decision computed by the PRODYN is the result of a rolling-horizon approach in which the performance is given by the estimated sum of the vertical queue length over all streams plus a

terminal cost associated with any remaining queues at the end of the horizon. The evolution of queues during the planning period is predicted by the queue state equation. The control strategy is the decision whether or not to switch from one stage to another, taking into account the predicted queue lengths and arrivals during the planning period. They tested three different flows (300, 400 and 500 vehicle/hour) using 5 seconds resolution time over a 75 seconds rolling-horizon period, and achieved a reduction in delay of about 10 per cent by comparison with optimal fixed-time control.

6) Vincent and Pierce's MOVA

After Vincent and Young (1986), the principles of MOVA (Microprocessor Optimised Vehicle Actuation) were described by Vincent and Pierce (1988). The aim of their research was to avoid the tendency of gap-seeking control systems (i.e. the UK System D Vehicle Actuated method: Department of Transport, 1984) to extend green times unnecessarily when traffic is flowing at less than full saturation flow. Based on Miller's self-optimising method, MOVA was designed for traffic-responsive signal control at isolated junctions with a time resolution of 0.5 seconds.

MOVA plans a sequence of decisions during the green period based on traffic flow and queue information derived from the vehicle detectors. The first decision is to estimate the minimum green time needed to clear the queue between the detector and the stop-line. After the full minimum green period, MOVA checks whether or not each stream currently receiving green has reached end-of-saturation before a change to the next stage is considered; until this decision is made, MOVA extends the current stage, subject only to the maximum green. Once the durations of end-of-saturation flow have been observed for all relevant streams, MOVA estimates the benefits and disbenefits of continuing the current stage green. Like Miller's method, MOVA considers extension of the current stage according to the objective of minimising the total rate of vehicle delay. The MOVA method is implemented in proprietary software. Unfortunately, neither the current literature nor the MOVA software documentation contain details of the theory that undertakes the method, so its performance cannot be assessed by simulation. For this reason, no comparison with MOVA is included in this study.

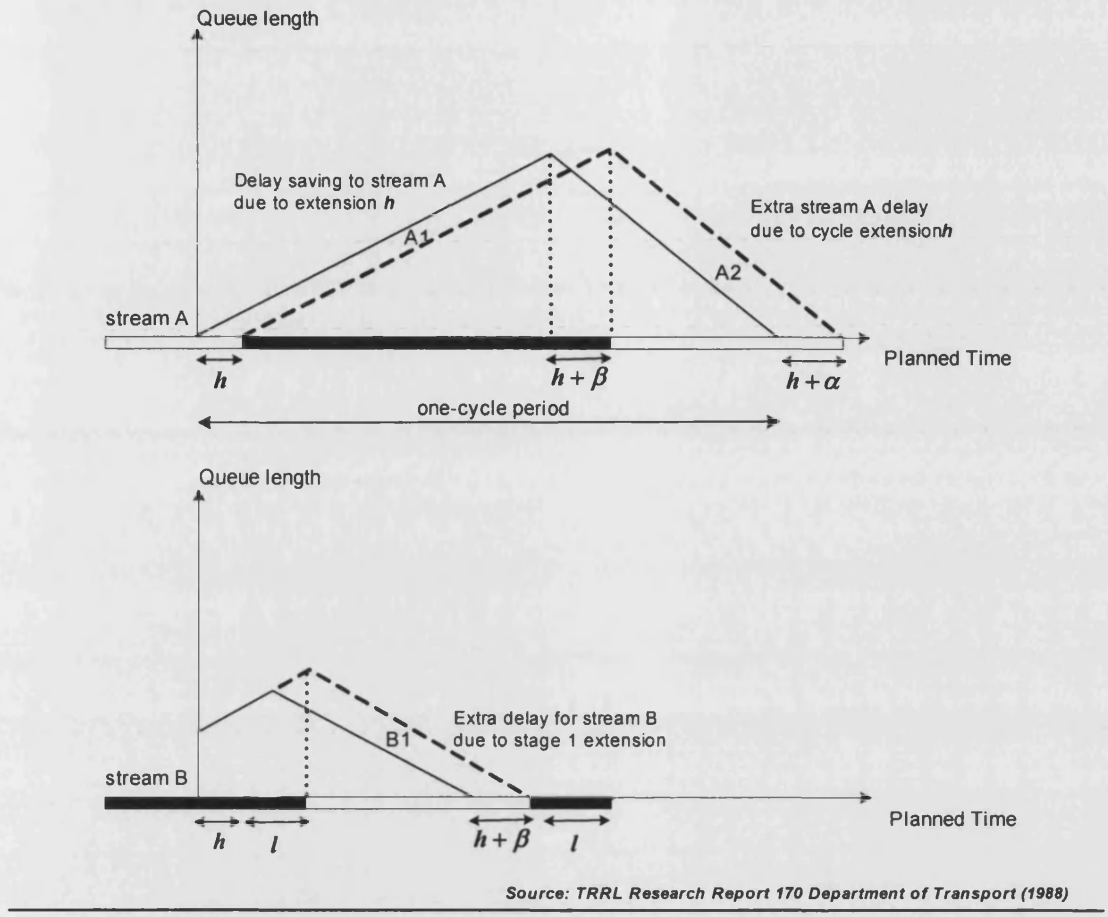


Figure 2.5 Delay changes as used by the MOVA optimiser

Figure 2.5 shows the components of the delay changes caused by extending the current green by h seconds, for a simple two-stream and two-stage junction. MOVA estimates the delay saved by stream A traffic using the green extension (depicted by area A1) and the delay caused by the traffic waiting for the next green (depicted by area A2) due to the cycle extension. As well as the delay caused by cross-road traffic waiting extra time of h seconds (depicted by area B1) is estimated. Finally, the net delay is calculated by summing A1, A2 and B1 delays. As can be seen in Figure 2.5, if the green is extended by h seconds, the start of the next green is deferred by $h+\beta$ seconds and respectively cycle time is extended by $h+\alpha$ seconds, but Miller assumes that α , β =0. In addition, MOVA has special features to manage situations when one or more streams to a junction are left with a significant queue at the end of the green and are judged to be oversaturated. MOVA automatically recognises oversaturation and switches from the normal delay-and-stops minimisation to a capacity maximising process to clear the congested approaches as quickly as possible.

According to the results of field tests, MOVA reduces vehicle delay by an average of 13 per cent compared with the system D vehicle actuated (VA) control.

7) Heydecker's group-based optimisation

Heydecker (1990) developed a systematic approach to traffic-responsive signal control which can be applied to a wide range of junction configurations. He used a group-based (or phase-based) optimisation method (Heydecker and Dudgeon, 1987) within the rolling-horizon approach. This approach is one of the bi-level optimisations of estimated control performance. At the upper level, the order in which control decisions are planned to be implemented is specified, whilst at the lower level the timing of these events is optimised for each of the specified orders. The combination of ordering and optimised timings which gives rise to the best estimated control performance is then implemented for the rollforward period.

The lookahead period is the same for all streams and includes a red and a green interval for each stream. The objective function is an estimate of the rate of delay incurred by traffic due to control decisions planned for the lookahead period, and has three distinct components of delay: detection period delay, prediction period delay, and terminal cost functions. In this respect, it is similar to that of Miller's method and MOVA. The use of terminal cost functions is based on the findings of Robertson and Bretherton (1974).

8) Heydecker and Boardman's DYPIC

Heydecker and Boardman (1999) extracted relevant information from analysis of video images, and developed backward dynamic programming formulations using each of 5 seconds and 0.5 seconds resolutions for decisions. They showed that the computing power of recent computers is adequate to solve these formulations within their respective time frames and that the additional delays caused by non-zero initial queues are all incurred during the early part of the planning period. This has the important consequence of limiting the excess costs associated with non-zero queues at the end of the planning period.

2.5 DISCUSSION

In this chapter, various optimisation methods were reviewed for fixed-time and traffic-responsive signal control. The main difference between these two control methods is the kind of traffic flow data that are used for optimisation: respectively historic flow data or on-line detector data. Here, the historic flow data is the pre-surveyed mean arrival rate, whilst the on-line detector data is the time dependent flow information provided from vehicle detectors. Hence, the performance of the fixed-time signal control can only be evaluated after the signal timings have been specified. In contrast, the performance of control decisions in traffic-responsive signal control has to be evaluated on-line.

In fixed-time signal control, the phase-based optimisation is particularly useful for more complex junctions, where the optimal stage sequence and interstage structures are not apparent. The optimisation on phases rather than stages can give considerable benefits in terms of reduced cycle time and delay or increased reserve capacity over existing methods (Silcock, 1990). In phase-based optimisation, the stage sequences can be generated by a branch-and-bound approach, but not all sequences satisfy the constraints and are necessarily acceptable in practice; consideration of safety and familiarity to road-users also play a role (Allsop, 1992).

Various optimising methods in traffic-responsive signal control have been developed by many authors, and several of them are currently used in practice at signalised junctions. In the past, theoretical methods developed for the problem of this optimisation have depended on data concerning arrivals for a considerable time into the future. This simplifies analysis and permits good performance to be achieved in theory, but it is often impractical. On the other hand, practical traffic-responsive methods (non-optimising and optimising methods), which use real-time data that are available from vehicle detectors have been developed in various ways. The common aim of these methods is to take advantage of short-term variations in the arrival patterns of the traffic in order to minimise delays, and hence to improve the quality of the traffic flow. As noted, the non-optimising methods, such as the VA control, do not use an objective function, but control decisions are made according to certain rules of operation. In contrast, the optimising methods use a specific objective function within their optimisation framework, in which control decisions are made on the basis of the estimates of delay incurred by traffic due to planned signal timings for the lookahead period.

Miller's (1963) self-optimising method used varying lookahead periods. This period was used untill the end of the next planned green for each stream, and hence it differed between them. Control decisions were made by the current stage extension test, in which the objective function estimated the difference of delay between the streams currently on green and red. This method assumes a

constant mean travel time from the detector to the stop-line. In contrast, Gartner's (1983) OPAC used the fixed lookahead period of about average cycle time, and hence it was the same for all streams at the junction. Control decisions were made by three changes of stage in the lookahead period, in which the objective function estimated the total delay incurred over all streams at the junction. The basic principle in common for both control methods is seeking the best control decision within a given lookahead period.

There are still some important features which remain to be investigated in a number of respects:

- A traffic model: interpretation of detector data (trajectory model) and estimation of operational performance (objective function).
- A dynamic optimiser: a systematic approach of defining lookahead periods with respect to varying signal timing plans, and making control decisions.

In Chapter 4, a novel traffic model is developed, and in Chapter 5 a systematic optimising method that includes the traffic model and the dynamic optimiser is presented and discussed.

3. TRAFFIC MODELS AND SIMULATION

3.1 INTRODUCTION

This chapter provides the necessary knowledge and general background of traffic models, which includes *car-following models* and *traffic flow models*. Some widely known traffic models are presented, and their capabilities and shortcomings are identified and discussed. In addition, two computer-based simulation packages, the TRANSYT (a mesoscopic model) and the SIGSIM (a microscopic model) are discussed in relation to traffic signal control.

Traffic models can be separated into two broad groups according to their scale of representation of traffic: *microscopic,* and *macroscopic* or *mesoscopic* models. In microscopic models, the movement of individual vehicles and the interaction between them is represented explicitly (by the representation of single vehicles) at a high level of detail, and are known as *car-following models.* By contrast, macroscopic models represent traffic at a more aggregate level (by the scalar value of flow rate), and are known as *traffic flow models.* We tend to choose microscopic approaches when details are important either in the modelling process or in the evaluation. On the other hand, we can choose macroscopic approaches when we are more interested in the large scale behaviour of the system.

An intermediate between microscopic and macroscopic level of analysis can be described as mesoscopic models. It is appropriate to call this a combined analysis method rather than a model because this kind of analysis comprises complementary combinations of macroscopic models; for example, where a macroscopic analysis is preferred for computational or other reasons but is known to cause error by simplicity, suitable corrections can be made by adjusting the results. The mesoscopic analysis adopts a simplified state description with simplified dynamics, but is distinguished from macroscopic ones by the inclusion of some correction for its simplification.

An objective of studying traffic models is to have a description of traffic flow on a set of roads or a network to make possible detailed integration of control strategies so as to reduce delays or improve flows. These models should be able to represent current stream conditions, and to predict the evolution of traffic into the future from a specified set of known initial conditions. Hence, the real process of traffic flows can be predicted by means of mathematical models without observing stream conditions. For example, if a certain number of vehicles had to traverse a given road section, then their expected travel speed and time can be estimated beforehand. Also, in more detail, time

dependent motion for individual vehicles can be estimated according to some model relationship, provided that sufficient initial data, such as position and speed, can be obtained for each of the vehicles.

3.2 CAR-FOLLOWING MODELS

The focus of car-following models is the relationship between the motion of two successive vehicles. Car-following models in a single lane assume that the motion of the first vehicle (leading vehicle) is determined exogenously, such as by traffic signal indications or traffic conditions at the downstream end, and the motion of the second vehicle is affected by the behaviour of the leading vehicle according to some vehicle following relationship. Car-following models regard the relationship between a set of successive vehicles by supposing that the following vehicle is influenced by the behaviour of the preceding vehicle. Similarly, the driver of the following vehicle attempts to maximise his speeds while maintaining a safe following distance. We can use a model of this form to represent the motion of a group of vehicles by supposing that the motion of the leading vehicle is known and then applying the model to calculate that of each successive vehicle.

3.2.1 General form of car-following models

The general form of car-following models is based on a *stimulus-response* relationship; with respect to the current speed and position of the leading vehicle, the following driver makes a decision (braking or acceleration) after his reaction time τ (e.g. Gipps used τ in general 2/3 seconds). The stimulus for the following vehicle can be taken as the speed difference between that vehicle and the one in front at time t, and the following sensitivity coefficient can be determined for the response at time $t+\tau$. The response (braking or acceleration) is dependent linearly on each of the stimulus and a sensitivity coefficient. This can be expressed in broad terms as

$$response(t+\tau) = sensitivity(t) \times stimulus(t) \tag{3.1}$$

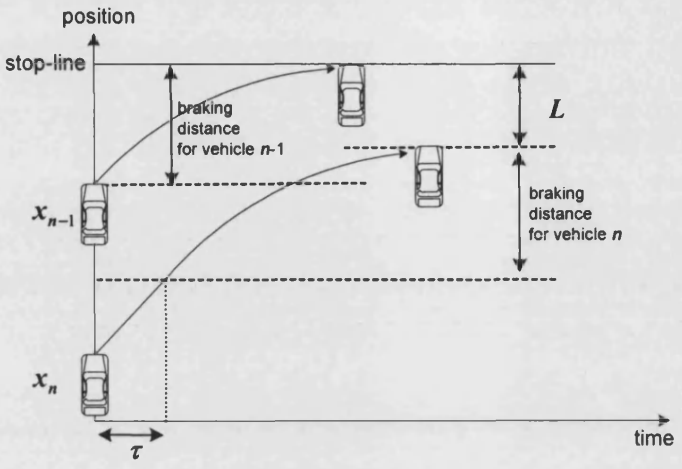


Figure 3.1 Car-following concepts

The notation that we adopt for car-following models is as follows.

Let

$x_n(t)$ be the position of vehicle n at time t,

$v_n(t)$ be the speed of vehicle n at time t,

τ be the duration of the time lag between stimulus and response,

L be the effective vehicle length including a safety margin,

$S_n(t)$ = $x_{n-1}(t) - x_n(t)$ is the space between a pair set of vehicles,

$a_n(t)$ be the acceleration rate at time t ($a_n(t) > 0$),

$b_n(t)$ be the constant braking rate at time t ($b_n(t) > 0$),

K be traffic density,

q be traffic flow.

If sudden braking is necessary for the leading vehicle, the following vehicle must have enough distance for a reaction and a braking period in order to avoid any possible collusion, which is described as a minimum safe spacing. The minimum safe spacing can be determined from

$$x_{n-1}(t) + \frac{v_{n-1}^2(t)}{2b_{n-1}} - L = x_n(t) + v_n(t)\tau + \frac{v_n^2(t+\tau)}{2b_n} \tag{3.2}$$

where $\frac{v_n^2(t+\tau)}{2b_n}$ is the braking distance of vehicle n. From the equation (3.2), we can find safe spacing as

$$x_{n-1}(t) - x_n(t) = v_n(t)\tau + \frac{v_n^2(t+\tau)}{2b} - \frac{v_{n-1}^2(t)}{2b} + L \tag{3.3}$$

If both vehicles have the same braking distance, and there is no speed change during the reaction time period τ, then $v_n(t) = v_n(t+\tau)$ and $\frac{v_{n-1}^2(t)}{2b} - \frac{v_n^2(t+\tau)}{2b} = 0$. Thus, (3.3) becomes

$$x_{n-1}(t) - x_n(t) = v_n(t+\tau)\tau + L \tag{3.4}$$

Differentiating the equation (3.4) with respect to time t, and then expressing it in terms of an acceleration will give

$$a_n(t+\tau) = \tau^{-1}[v_{n-1}(t) - v_n(t)] \tag{3.5}$$

The simplest form of this kind of car-following model has a response that depends linearly on the stimulus and has a sensitivity that is constant. In particular, we suppose that the sensitivity is independent of both speed and spacing, and is the same for acceleration as it is for braking.

Chandler et al (1958) generalised the equation (3.5). This linear model is simple and straightforward, but it does not lead to a reasonable description of traffic flow because the linear car-following model specifies an acceleration response which is independent of vehicle spacing. The generalised linear car-following model is given by

$$a_n(t+\tau) = \alpha[v_{n-1}(t) - v_n(t)] \tag{3.6}$$

where α is a sensitivity coefficient, $\alpha = q_m$ (q_m is the maximum flow capacity).

Gazis, Herman and Potts (1959) suggested a more elaborate car-following equation based on Greenberg's (1959) traffic flow model, which has the same stimulus of speed difference as the equation (3.6); however, a sensitivity is not a constant but is inversely proportional to vehicle

spacing, so that it is greater for smaller spacing and diminishes for larger ones. The speed-density relationship in this model is bounded only at high density traffic condition, so that it gives unbounded estimate of speed at low density. Therefore, Greenberg truncated speed at a certain density. The high density car-following model is given by

$$a_n(t+\tau) = \frac{\alpha_1}{[x_{n-1}(t) - x_n(t)]}[v_{n-1}(t) - v_n(t)] \tag{3.7}$$

where α_1 is a sensitivity coefficient, $\alpha_1 = v_m$ (v_m is the maximum speed at q_m).

Edie (1961) investigated car-following models in low density traffic condition at which maximum flow is achieved. This is same as Underwood's (1961) speed-density relationship for low density traffic condition. Edie proposed use of Underwood's speed-density relations for low density traffic condition in conjunction with Greenberg's relations for high density traffic condition, with a transition at the point of maximum flow. Starting from the low density relations of Underwood, speed and flow are continuous at the transition through the maximum density until the jam density. Edie's car-following model for low density is given by

$$a_n(t+\tau) = \frac{\alpha_2 v_n(t+\tau)}{[x_{n-1}(t) - x_n(t)]^2}[v_{n-1}(t) - v_n(t)] \tag{3.8}$$

where α_2 is a sensitivity coefficient, $\alpha_2 = v_m$ (v_m is the maximum speed at q_m).

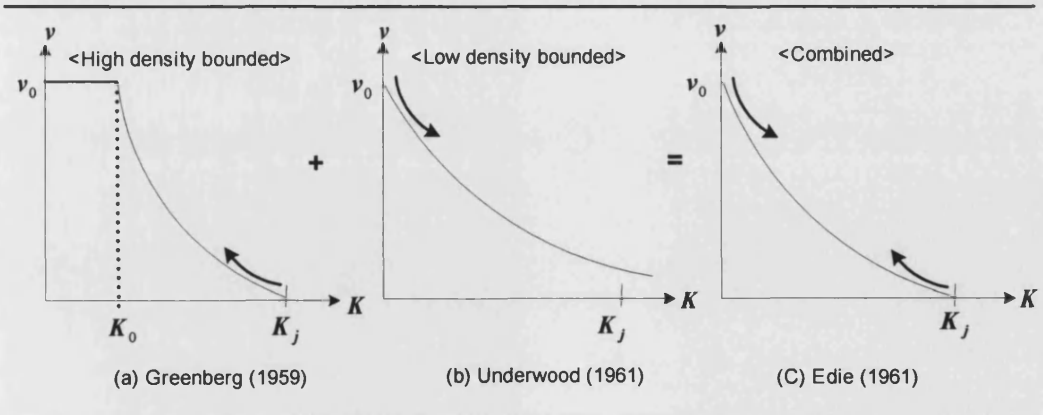


Figure 3.2 Speed-density curves (Greenberg, Underwood and Eddie)

All the above mentioned car-following models are included in a general non-linear model proposed by Gazis, Herman and Rothery (1961), in which the sensitivity is a generalised function of the speed of the following car and the spacing between leading and following cars. The generalised form of this model is given by

$$a_n(t+\tau)=\frac{\alpha_3[v_n(t+\tau)]^k}{[x_{n-1}(t)-x_n(t)]^l}[v_{n-1}(t)-v_n(t)] \tag{3.9}$$

where α_3 is a sensitivity constant to be determined experimentally with parameters k and l. For data taken by car-following runs in the Lincoln tunnel of New York City, Gazis et al found the best correlation for values k=1 and l=2. It is interesting to note that May and Keller (1967) investigated a range of possibilities, and found that the best fit for motorway data was achieved with values k=0.8, l=2.8, and for traffic in tunnels with values k=0.6, l=2.1.

3.2.2 Gipps' car-following model

Gipps' car-following model (1981) is a microscopic model. It is discrete in time but continuous in space. Gipps developed a car-following model based on the assumption that the driver of each vehicle sets limits to his desired braking and acceleration, in which the speed is chosen the minimum of v_n^a and v_n^b, which are calculated respectively according to acceleration and braking rules. The acceleration of the following vehicle assumes that a driver chooses the speed without exceeding the free-flow speed and its acceleration will never be greater than the specified maximum. Besides this, the braking is chosen on the assumption of maintaining the minimum safety spacing and pre-estimated maximum braking rate of the leading vehicle. If a vehicle brakes as hard as its driver desires, the following vehicle should be able to react and stop at a safe distance. In this model, vehicle positions and speeds are updated every 2/3 seconds.

For the acceleration speed v_n^a, two limits are involved, maximum acceleration rate and free-flow speed. The vehicle should not exceed its free-flow speed and the acceleration should first increase with speed as engine torque increases and then decrease to zero as the vehicle approaches the free-flow speed. These two limits are embodied in the following formula:

$$v_n^a=v_n(t)+2.5a_n\tau(1-v_n(t)/v_{0n})\sqrt{0.025+v_n(t)/v_{0n}} \tag{3.10}$$

where

$v_n(t)$ is the speed of vehicle n at time t,

a_n is the maximum acceleration that the driver of vehicle n wishes to undertake ($a_n > 0$),

v_{0n} is the free-flow (desired) speed at which the driver of vehicle n wishes to undertake.

From Gipps' acceleration speed equation (3.10), we can see why he took the acceleration parameter is 0.025 and how it constraints speed. The maximum acceleration value of a can be found following Spiropoulou (2003):

Let $a = \frac{v_n^a - v_n(t)}{\tau}$, then the equation (3.10) can be expressed as

$$a = 2.5a_n(1 - v_n(t)/v_{0n})\sqrt{(0.025 + v_n(t)/v_{0n})} \tag{3.11}$$

From (3.11), the maximum value of a can be achieved when $dv/dt = 0$, which gives

$$v_n(t) = \frac{1 - 0.025}{3} v_{0n} \tag{3.12}$$

From (3.12), we can get $v_n^*(t) = 0.325v_{0n}$, which means that if the speed proportion is $v_n^*(t)/v_{0n} = 0.325$, where the acceleration reaches its maximum. Likewise, applying (3.12) into (3.11), we can get $a = 0.998599a_n$, which is almost as high as the maximum acceleration parameters a_n. Thus, the acceleration of the following vehicle never exceeds that value.

For the braking speed v_n^b, Gipps adopted the parameter $\hat{b}$, which is the maximum braking rate that the driver of the vehicle n supposes for vehicle n-1 will not have the braking rate greater than that. Thus, the braking speed is

$$v_n^b = b_n\tau + \sqrt{(b_n\tau)^2 - b_n[2(x_{n-1}(t) - S_{n-1} - x_n(t)) - v_n(t)\tau - v_{n-1}^2(t)/\hat{b}]} \tag{3.13}$$

where

$x_n(t)$ is the location of the front of vehicle n at time t,

S_n is the minimum spacing of vehicle n, that is the physical length plus a safety margin,

b_n is the most severe braking that the driver of vehicle n wishes to undertake ($b_n < 0$),

$\hat{b}$ is the estimate of b_{n-1} since the driver of vehicle n cannot know this value by direct observation (set $\hat{b} > b_n$).

The speed at time $t+\tau$ is taken as the minimum of (3.10) and (3.13), so that

$$v_n(t+\tau) = \text{Min}(v_n^a, v_n^b) \tag{3.14}$$

Finally, the position updating equation is formed by assuming uniform acceleration between times t and $t+\tau$:

$$x_n(t+\tau) = x_n(t) + \frac{1}{2}[v_n(t) + v_n(t+\tau)]\tau \tag{3.15}$$

Gipps' car-following model is a detailed microscopic model. It is discrete in time but continuous in space. Thus, for the simulation, the speed and position of each vehicle is estimated for each time-step. This model was adopted in the microscopic computer simulation package SIGSIM (Silcock, 1993; Law and Crosta, 1999; Sha'Aban, 2003). More detail about the SIGSIM is presented in Section 3.6.

In order to apply this model at signal controlled junctions, a 'phantom' vehicle concept can be introduced to make the vehicle stop at the downstream stop-line; the phantom vehicle has zero speed and zero space so its position is at the stop-line of the junction. When the signal indication turns to red, this notional vehicle is placed in such a position and with such characteristics, in which the first vehicle approaching from the upstream can come to a stop at the downstream stop-line. The phantom vehicle is only operated if the signal turns to red.

3.2.3 Cellular Automaton (CA) model

Following Wolfram (1986), Nagel and Schreckenberg (1992) proposed the Cellular Automaton model (CA model), and implemented it for simulation of traffic flow on German streets and motorways. They found that this model showed reasonable representation on urban streets but not on motorways (Nagel and Schreckenberg, 1992).

The CA model is a simple microscopic single-vehicle simulation model in discrete time and discrete space. This model uses an integer value for speed (cell speed), which is considered as an important variable to evaluate performance. It is also called a particle hopping model because the vehicle 'hops' between cells according to its speed. The vehicle moves along a link, which is divided into cells, each one of which can be either empty or occupied by one vehicle. For example, if a vehicle speed is currently 2 at time t, then that vehicle moves 2 cells during one time-step.

Let

S be the length of a vehicle including a safety margin (S is equal to the length of cell),

v_0 be the desired speed (free-flow speed) of the link,

Δt be the time-step ($\Delta t \geq S / v_0$),

$v_n(t)$ be the cell speed of the vehicle n at time t,

$x_n(t)$ be the location of vehicle n at time t.

In time-step, a vehicle moves according to the current cell speed $v_n(t)$, then the new cell speed for time $t+\Delta t$ is calculated and updated based on the available number of empty cells, which is called the *gap*. The cell speed at time $t+\Delta t$ is determined explicitly by the gap and maximum speed limit. If the number of empty cell is greater than the current cell speed, then increase the speed by one; otherwise, the cell speed has to be less than or equal to the gap.

The principle of the Cellular Automaton model can be represented by following rules:

Step0 Find the number of empty cells (= *gap*) ahead at time t,
$gap = x_{n-1}(t) - x_n(t) - 1$.

Step1 If $v_n(t) < gap$ and $v_n(t) < v_0$ (enough spaces), then accelerate by one: $v_n(t) = v_n(t) + 1$.

Step2 If $v_n(t) > gap$ (too fast), then slow down to $v_n(t) = gap$.

Step3 Randomisation: after the above steps, if the speed is larger than zero, then with probability p, reduce the speed by one. With the introducing this rule, several values for the saturation flow can be accomplished, but this rule has no theoretical background and is introduced quite heuristically (Wu and Brilon, 1999).

Step4 Propagation: each vehicle moves with the cell speed $x_n(t+\Delta t) = x_n(t) + v_n(t)$.

In step 3, if the probability p is set to equal zero, then the model becomes deterministic, otherwise, with non-zero probability p the model becomes stochastic.

In the Cellular Automaton model, the length of cell is equal to a vehicle length plus a safety margin. If we suppose that the length of the cell is equal to 7.0 m in which one vehicle can be accommodated, and the desired speed of a link is equal to 14 m/sec, with time-step of $\Delta t = 1$ seconds the maximum cell speed of the vehicle can take is 2. This means that the time-step Δt has to be greater than or equal to the free-flow travel time for one cell, depending on the needs of the simulation. This model is a relatively simple microscopic model. It is discrete in time and also discrete in space. The way in which vehicles progress is estimated by searching the gap in front of them according to which they increase or decrease their speed. In this model, the acceleration cannot be more than one unit of speed per unit time, whereas when braking a vehicle can decelerate from the free-flow speed to zero speed in one time-step, which can be unrealistic. Vehicles only slow down when the vehicle in front is close and they only stop when the vehicle in front comes to a complete stop.

In order to apply this model at signal controlled junctions, a phantom vehicle with zero cell speed can be introduced in the cell that is downstream of the stop-line. So when the signal indication turns to red, this notional vehicle operation with zero cell speed can make the first vehicle comes to a complete stop at the downstream stop-line.

3.3 TRAFFIC FLOW MODELS

3.3.1 General form of traffic flow models (q-K relations)

Traffic flow models are concerned with finding general relations between the three fundamental variables of traffic: flow q, speed v and density K. These describe the average behaviour of traffic flow over different locations and different observation periods. This is a macroscopic approach that can provide an aggregate level of performance in the large-scale behaviour of system in which the behaviour of individual vehicles cannot be distinguished. The fundamental relationship between these three variables $q = vK$ is established and applied to describe traffic streams, including the propagation of shock waves and other phenomena in traffic.

Under the assumption of uniform flow (constant spacing and speed), the density is given in terms of number of vehicles per length of road, so that average spacing is $S = 1/K$. The number of vehicles

counted at the point of observation divided by the total observation time is defined as flow, and the reciprocal of flow can be interpreted as headway $h = 1/q$. If two vehicles are travelling at spacing S and identical speed v, the headway between these vehicles is defined as $h = S/v$, where v is the space mean speed (harmonic mean of spot speeds). Substituting each relation, the fundamental equation of traffic flow is

$$q = vK \tag{3.16}$$

The fundamental hypothesis of the first order traffic flow theory is that the flow is the function of density $q = f(K)$, in which the traffic stream is treated as a one dimensional compressible fluid.

3.3.2 L-W-R model (Wave model)

Lighthill and Whitham (1955), and Richards (1956) pioneered traffic flow models, which is known as L-W-R theory. They described a theory of one dimensional wave motion on the basis of continuum fluid flow. This theory has two basic assumptions: a) there is a one-to-one relationship between speed and density; b) traffic is conserved. The first principle has raised many objections in the literature partly because of contradictory measurements; it has been observed that for the same values of density, many values of speed can be measured. The relation between flow and density, or between speed and density, is not valid under time-dependent condition but valid only at equilibrium. The second principle is generally acceptable and can be expressed the conservation rule (3.17), which implies that in any traffic system input is equal to output plus storage.

The conservation condition can be expressed in the partial differential equation:

$$\frac{\partial q}{\partial x} + \frac{\partial K}{\partial t} = 0 \tag{3.17}$$

where flow q and density K are two fundamental dependant variables, and time t and distance x are independent ones along the link.

This equation has the same form as in fluid flow. As we can see above, this equation is related to two dependent variables, density and flow. Considering the fundamental relationship $q = vK$, speed

can be expressed as a function of density, so that $v = f(K)$. Then the conservation equation (3.17) can be rewritten as:

$$\frac{\partial(vK)}{\partial x} + \frac{\partial K}{\partial t} \Rightarrow \frac{\partial[Kf(K)]}{\partial x} + \frac{\partial K}{\partial t} = 0 \tag{3.18}$$

by using the produce and chain rule, (3.18) becomes

$$\frac{\partial K}{\partial x} f(K) + K[\frac{df(K)}{dK} \frac{\partial K}{\partial x}] + \frac{\partial K}{\partial t} = 0$$

or

$$[f(K) + K\frac{df(K)}{dK}]\frac{\partial K}{\partial x} + \frac{\partial K}{\partial t} = 0 \tag{3.19}$$

where $f(K)$ can be any speed-density function, so that no further assumptions are needed to keep the results general.

The density K is constant along a family of curves called waves; a wave represents the motion of a change in flow and density along the roadways. From the relationship (3.19), we can express that $\frac{dK}{dt} = \frac{dx}{dt}\frac{\partial K}{\partial x} + \frac{\partial K}{\partial t} = 0$ for constant K. From the relationship in (3.17 and 3.19), the speed of waves is given by

$$\frac{dx}{dt} = f(K) + K[f'(K)] = \frac{dq}{dK} \tag{3.20}$$

According to the present model, density determines speed and hence flow; the speed of a wave depends only on the traffic density, which is constant on a wave, so that the speed of a wave is itself constant. From the equation (3.20), we can get the wave speed $v_w(K)$, that is

$$v_w(K) = \frac{dq}{dK} = v(K) + K\frac{dv}{dK} \tag{3.21}$$

where $v(K) = q/K$, which is the average speed (space mean speed) and $dv/dk < 0$ (in general speed decreases with increasing density).

At time $t = 0$, if we know the position $x_K(0)$ of a region of density K, then we can locate it at later times according to the linear formula, which is given by

$$x_K(t) = x_K(0) + v_w(K)t \tag{3.22}$$

As seen in Figure 3.3, any points on the $q - K$ curve can be used to represent the flow conditions. The slope at the tangent of the flow density curve is equal to the wave speed (3.21). The wave peed is always slower than average vehicle speed $v_w(K) \leq v(K)$ when $K>0$ and equal to $v(K)$ only when $K=0$. The sign of wave speed $v_w(K)$ can be either positive or negative, depending on the value of K. If this speed is positive, that means that a region of density K is moving in the direction of downstream because the downstream traffic condition is uncongested. However, when this speed is negative, which means that a region of density K is moving in the direction of upstream because the downstream traffic condition is congested.

3.3.2.1 Shock wave model

Using the L-W-R wave model, a shock wave is said to exit where two different traffic conditions meet. If the sign of the shock wave speed is positive, the shock wave moves to the downstream direction; if it is negative, the shock wave moves in the upstream direction; if it is zero, the shock wave is stationary. The speed of the shock wave can be given by the slope of the chord connecting between two stream conditions. The speed of shock wave v_{sw} between downstream (D) and upstream (U) is given by

$$v_{sw} = \frac{q_D - q_U}{K_D - K_U} \tag{3.23}$$

For example, Suppose that the linear speed density function of Greenshields (1934) is applied to the wave speed equation (3.22). From Greenshields' speed-density function of $f(K) = v_0(1 - K/K_j)$, we can derive $f'(K) = -v_0 / K_j$, then the equation (3.21) becomes

$$v_w(K) = v_0(1 - 2\frac{K}{K_j}) \tag{3.24}$$

Using Greenshields' (1935) speed-density function, we can get the maximum flow density $K_m = K_j / 2$, and its maximum flow $q_m = v_0(1 - K_m / K_j)K_m = v_0 K_j / 4$.

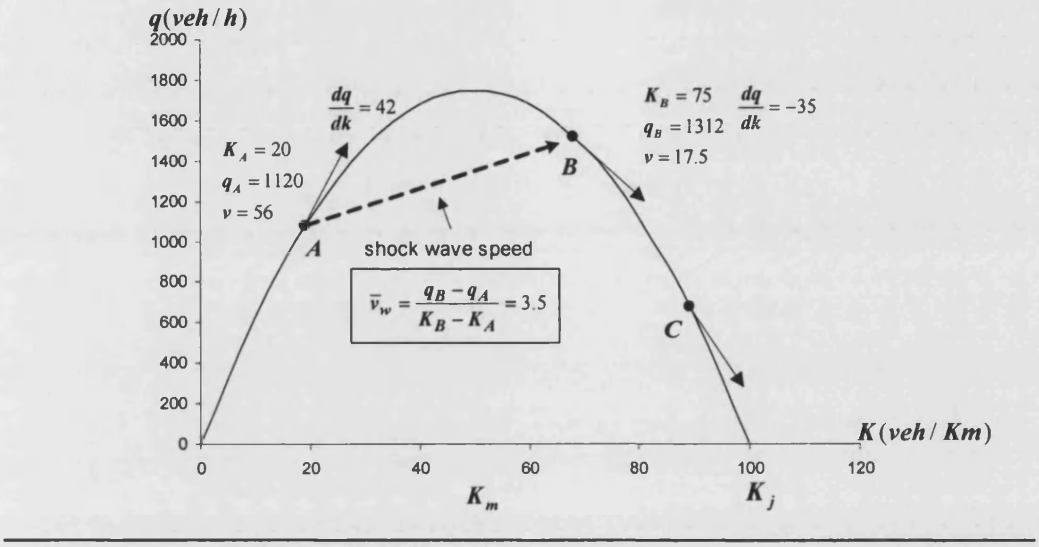


Figure 3.3 Mean speed, wave speed and shockwave speed in $q - K$ curve

As shown in Figure 3.3, suppose that state B decides the downstream traffic condition, and state A decides the upstream traffic condition. When the free-flow speed v_0 = 70 Km/h and the jam density K_j = 100 veh/km, under these conditions we can get K_m = 50 veh/km and q_m = 1750 veh/hour. From this example, we can say that the downstream (state B) traffic condition is more congested than the upstream (state A). State B corresponds to the density of K_B = 75 veh/km, flow of q_B = 1312 veh/hour, and the average speed of v_B = 17.5 km/h. In contrast, state A corresponds to the density of K_A = 20 veh/km, flow q_A = 1120 veh/hour, and the average speed of v_A = 56 km/h. When state A meets state B, the shock wave speed v_{sw} = 3.5 km/h (3.23).

The speed of the shock wave is less than the speed of the vehicles (average speed) on either side. In state B traffics are travelling at 17.5 km/h, whilst approaching vehicles in state A are travelling at 56 km/h, but the shock wave between the two platoons is 3.5 km/h. Even though the state B platoon forced the traffic to slow down, the flow increased from 1120 to 1312 veh/h. In certain cases, slowing the traffic via the traffic control system may be a good way of increasing the flow, but we can suppose that this consequence goes unnoticed by the drivers. Additionally and briefly, if state C

meets state A, the shock wave speed is negative, which means that the congested region extends upstream over time (the queue is forming to the upstream direction).

3.3.3 Daganzo's Cell Transmission model

Daganzo (1994) developed a Cell Transmission model (CTM model) based on a discrete approximation of the L-W-R model of traffic flow. The CTM model represents the movement of traffic over discrete time and discrete space, including transient phenomena such as building, propagation and dissipation of queues. The movement of traffic is based on the minimum amount of traffic that can be sent from the upstream cell and the amount of traffic that can be received in the current cell.

Let

v_0	be the free-flow speed,
Δt	be the time-step,
S	be the length of the vehicle ($S = 1/K_j$),
$n_{i-1}(t)$	be the number of vehicles in cell i-1 at time t,
$N_i(t)$	be the maximum occupancy of the cell i at time t ($N_i(t) = v_0 \Delta t / S$),
$Q_i(t)$	be the maximum flow from cell i-1 to the current downstream cell i at time t ($Q_i(t) \le N_i(t)/2$),
$N_i(t) - n_i(t)$	be the amount of available empty space in cell i at time t.

This model uses homogeneous cell spaces. The cell length is chosen to be equal to the free-flow travel distance during a time-step Δt. This means that under light traffic conditions all vehicles in a cell can move to the next cell in one time-step. Thus, the evolution obeys the rule that the number of vehicles in cell i at time $t + \Delta t$ is equal to the number of vehicles in cell i -1 at time t : $n_i(t + \Delta t) = n_{i-1}(t)$.

With known free-flow speed v_0, the maximum occupancy of a cell can be calculated as $N_i(t) = v_0 \Delta t / S$ (Daganzo defines S = 6.5m in his model). By setting the maximum occupancy of the cell, the cell length is defined as $N_i(t) \times S$. The only constraint applied in this model is the

maximum flow $Q_i(t)$. This quantity can not take a value higher than half of the cell occupancy per time step, so that $Q_i(t) \leq N_i(t)/2$.

The CTM simulation is based on the recursion where the amount of traffic in a cell at time $t+\Delta t$ is equal to the amount of traffic occupying in it at time t plus inflow $y_i(t)$ from the upstream cell minus the outflow $y_{i+1}(t)$ to the downstream cell, where current stream conditions of all cells are updated every time-step Δt. This can be expressed as:

$$n_i(t+\Delta t) = n_i(t) + y_i(t) - y_{i+1}(t) \tag{3.25}$$

where $y_i(t) = \text{Min}[n_{i-1}(t),\ Q_i(t),\ N_i(t) - n_i(t)]$. Hence, the movement of traffic is based on the minimum amount of traffic that can be sent from the previous cell, $n_{i-1}(t)$ and the amount of traffic that can be received in the current cell, $N_i(t) - n_i(t)$ subject to a maximum flow of $Q_i(t)$.

The Cell Transmission model is a straightforward macroscopic model. It is discrete in both time and space. The formula that defines the movement of the vehicles is based on the assumption that a driver will move when there is empty space in front of him. A shortcoming of using the CTM for simulating traffic is that vehicles can implicitly accelerate from zero to free-flow speed, or decelerate from free-flow speed to zero, in one time-step. This assumption keeps the model simple, but it is unrealistic; driver behaviour characteristics are not included in the formula.

In order to apply this model at signal controlled junctions, the maximum amount of flow $Q_i(t)$ has to be manipulated during the signal operation. When the signal indication turns to red, the maximum flow is set to zero for the cell at the stop-line. Thus, no flow is allowed to the downstream cell during the red period.

3.4 VERTICAL QUEUEING MODEL

The vertical queueing model is relatively simple, and hence widely adopted in traffic signal optimisation. Due to the simplicity of the vertical queueing concept, this model can be used in microscopic, macroscopic and mesoscopic simulation.

The vertical queueing model represents all vehicles (or flows) entering a link have the same travel time (free-flow travel time) to the downstream stop-line, at which vehicles join a queue vertically, not horizontally. It assumes that all queueing vehicles move with the same speed, and stop instantaneously. Once vehicles are held in a vertical queue, the departure time of the first vehicle is assumed to coincide with the start of the effective green time (Clayton, 1940; Webster, 1958; Allsop, 1970). Further following vehicles are assumed to depart the stop-line at equal headways (saturation departure time).

The start of the effective green time t_{ge} is calculated from the beginning of the indicated green time t_g (or the beginning of the red and amber, depending on the driving behaviour) with respect to the free-flow speed and the acceleration rate, that is given by

$$\begin{aligned} t_{ge} &= t_g + (\frac{v_0}{a} - \frac{\bar{X}_v}{v_0}) \\ &= t_g + \frac{v_0}{2a} \end{aligned} \tag{3.26}$$

where v_0 is the free-flow speed, a is the acceleration rate, and $\bar{X}_v$ is the position where the stopped vehicle regains the free-flow speed after taking a full acceleration ($\bar{X}_v = v_0^2 / 2a$).

In applications of microscopic simulation, following vehicles will depart the stop-line with a constant headway (saturation departure time), but in macroscopic simulation cumulative queueing traffic will dissipate with the saturation departure rate because it treats traffic flow rather than individual vehicles.

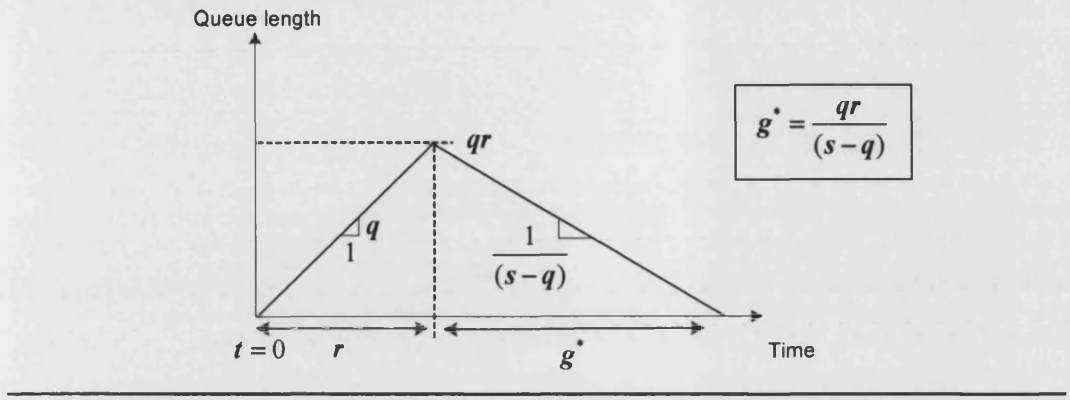


Figure 3.4 Queue formation and dissipation in the vertical queueing model

As seen in Figure 3.4, the cumulative queue length during the red period r is obtained by qr, and the queue clear (dispersion) time is calculated by $g^* = qr/(s-q)$, where q is the flow rate and s is the saturation departure rate. The total delay occurs during this time period $(r + g^*)$ is given by

$$D_V = \frac{sqr^2}{2(s-q)} \tag{3.27}$$

By contrast, in microscopic models, the delay is identified as the time difference between a queue departure time and the free-flow travel time of each individual vehicle. Further details about this model are discussed in Chapter 4, Section 4.5.1.1.

The basic concept of the vertical queueing model was adopted in the computer package TRANSYT (Robertson, 1969; Vincent et al, 1980; US DOT FHWA, 1991). This program is widely used to calculate and investigate timings for traffic signals in urban road networks. Further details about the TRANSYT traffic model are discussed in Section 3.5.

3.5 TRANSYT TRAFFIC MODEL

The principle of the TRANSYT traffic model was firstly described by Robertson (1969). The TRANSYT traffic model is a good example of the mesoscopic model, because it comprises two macroscopic models: the TRANSYT traffic model and the signal optimiser. The first one is the platoon dispersion model, which uses the vertical queueing concept (see Section 3.4), and the second

one is the random and oversaturated delay formula. TRANSYT is a method of finding the best fixed-time signal plans with which to coordinate the traffic signals in any networks of roads for which the average traffic flows are known. TRANSYT traffic model has been successfully used for optimising signalised junctions over 40 years.

This model is discrete in time, but there is no space element, which means that vehicles are either in a link or not. It does not provide any details of their position. The network is represented by 'nodes' and 'links'; each signalised junction is represented as node and they are inter-connected with links. This programme assumes vertical queueing at the stop-line in the sense that all vehicles take the same travel time between the nodes; no information about flows within the link is provided. In other words, TRANSYT simplifies the behaviour of traffic by assuming that vehicles in a link are undelayed until they reach the stop-line.

In addition, TRANSYT is a stage-based optimisation program including a numerical model of traffic movement in which platoon of vehicular movements between adjacent junctions are described using a platoon dispersion factor. During the optimisation, the common cycle time of the signals is divided into a number of equal intervals called time-steps (typically 1-3 seconds long). All calculations are made on the basis of the average values of traffic flow rates, turning movements and queues, which are expected to occur in each time-step of a typical cycle. The calculations of the behaviour of traffic are made by manipulating the following three types of cyclic flow profile:

1) IN profile: the arrival pattern at the downstream stop-line if the traffic were not impeded by the signals at the stop-line.

2) OUT profile: the departure pattern of traffic actually leaving the stop-line, it is usually equal to the GO pattern as long as there is a queue, after the queue has discharged, it is equal to the IN pattern for the duration of the effective green time.

3) GO profile: the pattern of traffic that would leave the stop-line if there were enough traffic to saturate the green.

The traffic flowing into a link is obtained by taking the appropriate fraction of the OUT profiles from the upstream links. Namely, the upstream OUT profile is taken into the downstream IN profile, and platoons of vehicles are partly dispersed by applying a form of exponential smoothing to the incoming traffic. The degree of smoothing is a function of the cruise time along the link. It is the average undelayed travel time for vehicles flowing from the upstream stop-line to the stop-line at the

exit end of the link. The IN profile at the downstream is calculated recursively from the upstream OUT profiles using the platoon dispersion equation (Robertson, 1969). For each time-step k, the arrival flow at the downstream stop-line (ignoring the presence of a queue) is found by the following recurrence equation:

$$q^{L}_{(k+t_F)} = Fq_k + (1-F)q^{L}_{(k+t_F-1)} \qquad (\text{for } k = 1,2,\ldots N) \tag{3.28}$$

where

q_k is the flow in the step k of the upstream OUT profile (the flow rate of the initial platoon),

F is the smoothing factor equation, $F = 1/(1+\beta t_F)$,

t_F is α times the mean cruise time (measured in steps) over the distance for which dispersion is being calculated,

q^{L}_{k} is the flow in the step k of the downstream IN profile, Robertson's platoon dispersion model supposes that each q_k contributes $F(1-F)^r q_k$ to each $q^{L}_{(k+t_F-1)}$ for $r = 0, 1, 2,\ldots,$

The parameters α and β can be calibrated, but current default values are α =0.80 and β =0.35 (Vincent, Mitchell and Robertson, 1980).

The number of vehicles (m_i) that is held at the stop-line during the interval i is given by

$$m_i = \text{Max}(0,\ m_{i-1} + q_i - s_i) \tag{3.29}$$

where

q_i is the number of vehicles arriving in interval i (given by the IN profile),

s_i is the maximum number of vehicles that are allowed to leave in an interval i (given by the GO profile).

The number of vehicles leaving from the stop-line during the interval i is given by $m_{i-1} + q_i - m_i$.

The performance index (PI) in this program is the weighted combination of the delay and stops on all the links in a network: a weighted sum of estimated rate of delay and number of stops per unit time. It can be expressed as

$$\mathrm{PI} = \sum_{i=1}^{L} (W_{Di} D_i + W_{Si} S_i) \tag{3.30}$$

where PI is the performance index of the network, D_i and S_i are respectively the rate of delay and number of stops on link i, W_{Di} and W_{Si} are respectively delay weighting and stop weighting on link i.

In the Equation (3.30), D_i consists of the sum of the rates of uniform delay, and random and oversaturation delay on link i. Similarly, S_i consists of the uniform number of stops and random number of stops on link i respectively. Refer to Robertson (Transyt: a traffic network study tool, 1969) and Vincent et al (User guide to TRANSYT version 8, 1980) for a detailed procedure of delay calculation.

TRANSYT is a fixed-time signal optimisation program. It comprises a traffic model to simulate flows, and a signal optimiser to estimate the PI. The aim of TRANSYT programme is to find good signal timings for either a single junction, or a road network under coordinated control by adjusting the signal timings repeatedly. The optimisation process adjust the signal timings and monitors, whether or not these adjustment reduce the PI. In such a way only changes in timings that reduce the PI are adopted.

3.6 SIGSIM SIMULATION MODEL

This section provides a general background of the SIGSIM simulation model, and puts more focus on the programming structure of handling event-based data. More details about using the SIGSIM traffic simulator, and how it is interfaced with a dynamic optimiser will be presented in Chapter 6.

SIGSIM is a microscopic simulation program that can be used to simulate the traffic behaviour of individual vehicles, either for an isolated signal controlled junction or for a network of signal controlled junctions under different signal control strategies. Individual vehicles are simulated using a car-following model developed by Gipps (1981) which calculates the speed and position of each vehicle on a lane according to each vehicle's individual characteristics and the position of the vehicle in front.

SIGSIM was originally developed to test real-time signal optimisation algorithms by the Transport Operation Group (TORG), Newcastle University and the Centre for Transport Studies (CTS), University College London. The original program has been enhanced to allow a network of vehicle-actuated controlled junctions to be modelled in a way that reflects a real urban signalised network. The latest version of SIGSIM (Silcock, 1993; Law and Crosta, 1999; Sha'Aban, 2003) which runs on PC-based programming environments (written in C) will be discussed in this section. The current version of SIGSIM can be used for simulating vehicle movement under five different signal control strategies so far (new control policies can also be added), which are:

a) Fixed-time control
b) Vehicle actuate system D control (gap-seeking method)
c) Bus-priority system D
d) Traffic responsive TORG control
e) Bus-priority TORG

An important feature of SIGSIM is that it has 'between runs variability'. This means that if SIGSIM is run twice with the same input data, it will produce the same results because vehicle arrivals are generated stochastically. In contrast, if the random number seed is changed, then it will produce different results representing the same junction and vehicle profiles. This feature is somewhat realistic and as consequences, it is important when comparing different signal control strategies or geometric changes that a number of runs are carried out to establish the degree of variation arising from random variations in traffic flow patterns. This is known as between runs variability. If the SIGSIM simulation is set to run more than once with the same input data, the final results will provide an estimate of mean performance that is qualified by the associated standard error to the number of runs with different random number seeds.

3.6.1 Event-based simulation

SIGSIM is an event-based simulation model. This means that, rather than updating the simulation at regular time intervals, the simulation is only updated at the occurrence of an event. The signal oriented method in SIGSIM runs according to the user defined time-step, which is called a 'controlscan event time' in SIGSIM. During this interval, if any new events are generated, such as vehicles crossing the detectors (event of detector_on) or traffic light changes (event of lightchange), the generated new events are stored into the current event list. After making the transitions corresponding to an event, all scheduled events are considered and the earliest of them is identified

for the next event. The simulation is continuously advanced by each event until the end of the simulation time, during which new events are continuously generated as a result of processing current ones. As with signal control strategies any new events we wish to obtain from the simulation can be added to the event list

In SIGSIM source code file <event.c>, the module of process_next_event function contains a list of all possible events in relation to their intended purpose of calculation. Respectively, schedule_event() function stores all generated events into the current list and schedule for the next event. Each time a new event is generated, it is placed in the current list in order of its scheduled time.

3.6.2 Vehicle updating and signal control policy

In SIGSIM, the behaviour of the vehicle is determined by Gipps' car-following model (1981), described in Section 3.2.2. All vehicles in the simulation at a given time have their position, speed and other characteristics, those are updated simultaneously once every updated period. This is set to 2/3 seconds, which is an appropriate estimate of driver reaction time for use in this model (Gipps', 1981).

Vehicles are generated independently for each lane on an entering link. According to the user defined in flow, each vehicle is generated based on the *shifted exponential distribution of headways* (see Section 4.5.2.2) with its own serial number. If the link is blocked by traffic, generated vehicles are stacked in their order of arrival and enter the simulation when the conditions allow. When a vehicle crosses a specific detector, it contains three components of information: detector number, time of detection and vehicle number. These are obtained in the source code file <traffic.c> in the module of switch_detector_on() function.

Signal control in SIGSIM is modelled separately from other simulation aspects. The code is contained in the source file <signalcontrol.c> and no variables are shared globally with other source files so that information about SIGSIM traffic must be passed (returned) as parameters and the signal timings are passed back to the simulation. If we wish to add a different signal control strategy, it can be modified in the module of signal_control function, in which the light status of each lane (or detector) with respect to the current simulation time can be obtained.

The signal control is based on the MCE 0141 controller specification (Department of Transport, 1984). In SIGSIM, the System D Vehicle Actuated (VA) control mode works by phases being demanded and extended via the activation of detectors. Each lane has three detectors and are placed at 12m, 26m and 40m from the downstream stop-line, and is 2m long. However, the position and number of detectors can be changed for evaluation purposes. When the fixed time control is in operation, the status of the detector is ignored, and all phases are set to being demanded and extended; this forces all phases to run to their maximum green time.

3.6.3 Junction geometry and delay

Vehicle behaviour close to an intersection will depend on the junction geometry and the traffic lights. SIGSIM takes junction geometry into account indirectly by reducing the desired speed of vehicles approaching an intersection by applying reduction variables to the desired speed of the vehicles, which is called a *speed reduction factor*. The Gipps' equations do not take into account the junction geometry and therefore the speed reduction factor has been introduced as affecting the driver's desired speed.

Vehicles are affected when they are within the zone of influence. This zone has been set to 50m before the stop-line and 5m beyond the stop-line. This is an important adjustment factor for controlling the saturation flow of a junction. The user can define the maximum speed reduction in input data file <junction.file>. For each lane, the following equation (Gaussian function) applies:

$$g = 1.0 - \alpha \exp\left(-\frac{x_1^2}{2l_1^2} - \frac{x_2^2}{2l_2^2} \right) \tag{3.31}$$

thus, the speed within the zone is given by $v = v_0 g$,

where

v_0 is the normal desired speed,

g is the factor applied to the normal desired speed at each position,

α is the speed reduction expressed as a proportion,

x_1 is the distance of an approaching vehicle from the stop-line x_s,

$x_1 = \min(x - x_s, 0)$, where x is the position of the vehicle at each time and x_s=0,

x_2 is the distance of a departing vehicle from the stop-line x_s,

$x_2 = \max(x - x_s, 0)$, where x is the position of the vehicle at each time and x_s=0,

x_1 is the distance of an approaching vehicle from the stop-line x_s,

l_1 is the zone of influence upstream of the stop-line (-50m),

l_2 is the zone of influence downstream of the stop-line (5m).

Various measures of performance can be used to compare each signal time plan with others. The mean delay for vehicles is the difference between the average journey time through the junction and the average journey time that would apply if the vehicle were not impeded by the signals. Thus, the product of mean delay and the arrival number of vehicles in a stream provides an estimate of the rate of delay, and has units of vehicles. So that total delay divided by the duration of the cycle time also can be used as an estimate of the mean rate of delay.

The final report of SIGSIM <finalreport.file> gives the rate of delay on each link and the total rate of delay for a whole junction. The reason for using the total rate of delay is that is additive over streams to give a measure of performance for the junction as a whole.

3.7 DISCUSSION

Various traffic simulation tools have been developed and used widely to investigate the performance of systems, and in particular their dynamic behaviour in detail. In practice, traffic behaviour is too complicated to describe analytically in detail because traffic has varying kinds of behaviour over time. In such complicated systems, simulation is a good tool to represent the process of real traffic explicitly. The idea behind simulation is to provide numerical results to the analyst for quantitative estimation of what is likely to happen. Thus, it should be able to provide estimates of the effects of any policy or measures of performance that the analyst wishes to evaluate with it.

These simulation models can be integrated with models of traffic signal optimisers. As discussed, the vertical concept has been used widely on the basis of the macroscopic level, in which the focus of modelling is put more on traffic (not individual vehicles) with relatively less consideration of the speed. On the other hand, microscopic traffic flow models can properly represent the motion of individual vehicles, but they are not ideally suitable for use with a dynamic signal optimisation because of their computational burden and the difficulty of achieving accurate estimates, in the presence of stochastic variation. In general, stochastic generation methods by working probabilistically are usually applied to microscopic simulations .

In the TRANSYT program, a macroscopic vertical queueing model is used in the TRANSYT traffic model. This computer package can only optimise fixed-time signal control under the given conditions of known traffic flows. Because the TRANSYT traffic model is mesoscopic, it is not suitable for a detector based traffic-responsive signal optimisation. By contrast, SIGSIM is an event-based microscopic simulation tool which works on the basis of detectors, in which any new events can be added and manipulated by the users. Hence, this kind of simulation tool is suitable for a detector based traffic-responsive signal optimisation. In fixed-time signal control mode, SIGSIM only evaluates performance in the role of a general simulation tool; it does not optimise signal timing. However, in System D vehicle actuated control mode, SIGSIM calculates signal timing based on the gap-seeking method, but has no specific objective function for control, so that this kind of vehicle actuated control is not normally optimal (Bell, Cowell and Heydecker, 1989).

In dynamic signal optimisation, the first important information we wish to obtain is the travel time of each individual vehicle, from the upstream detector to the downstream stop-line. It would be tedious and time consuming to describe all vehicular motion in detail at each time-step and position. Moreover, the motion of vehicles at signal controlled junctions is affected by the current status of traffic lights downstream, so more emphasis should be put on the motion descriptions as vehicles approach the stop-line. Based on the kinematics concept of physics, the novel trajectory model is developed in Chapter 4, and interfaced with the SIGSIM microscopic traffic simulator. In this way, the SIGSIM provides detector data to the trajectory model, and the trajectory model estimates all detailed motions from the detector to the stop-line.

4. A KINEMATIC CAR-FOLLOWING MODEL FOR SIGNALISED JUNCTIONS

4.1 INTRODUCTION

In this chapter, the motion of vehicles from the detector to the stop-line is formulated analytically as a function of the signal indication (start of the green time) with respect to the information provided by the detectors. Based on the one dimensional kinematic equations of physics, an alternative traffic model that can be applied to a signal timing optimisation is proposed in two different parts: *a trajectory of the leading vehicle* and *a trajectory of following vehicles*. Then the arrival and departure time of each vehicle at the stop-line can be estimated directly when it crosses the detector. Finally, the difference between the vertical queueing model and the novel traffic model in sensitivity of delay with respect to the variations in the start of green time is investigated and discussed.

The motion of vehicles approaching a signal controlled junction is affected by the current signal indication. The driver will make a decision to either brake, maintain current speed or accelerate according to whether the signal display at the junction is red, green or amber. Consequently, in traffic signal optimisation, the motion cannot be determined before a signal timing plan is proposed, so that the motion should be described as a function of the signal indication. In this way, the motion of the leading vehicle is determined by the signal indication and then successive following vehicles are estimated using the behaviour of the vehicles in front of them.

The vertical queueing model that is adopted widely in traffic signal control represents all vehicles have the same travel time before joining a queue. Thus, a queue is supposed to form vertically at the stop-line without occupying any space on the link. Due to the simplicity of this model, all vehicle motions are identical and unaffected by the signal display on their approach to the junction. In the vertical queueing model, the departure time of the leading vehicle is assumed to coincide with the start of the effective green time (Clayton, 1940; Webster, 1958; Allsop, 1970) and then departure times for the successive following vehicles are estimated in accordance with the saturation departure time (headways) at the stop-line. The delays and departure time estimates given by the vertical queueing model are not always realistic, because this model does not consider any braking motion until the stop-line. Namely, it is more focused on the queue delay rather than the approach delay (the consequences of this are discussed in Section 4.5). The TRANSYT (Robertson, 1969) uses vertical queueing concepts in fixed-time signal optimisation. Also, in traffic-responsive signal control, Miller (1963) and Gartner (1983) use constant mean travel time from the detector to the stop-line, and Bang

(1976) uses occupancy rate of the loop detector and the average speed to estimate arrival time at the stop-line.

In dynamic traffic signal control, the motion of each vehicle from the upstream detector to the downstream stop-line is needed for full interpretation of the detector output and for performance evaluation. It would be tedious and time consuming to describe the motion of each vehicle in detail at each time-step. Hence, the concepts of kinematics of physics are applied to derive *a Kinematic Car-following model at Signalised junctions* (KCS traffic model). According to this model, arrival times and departure times of all detected vehicles at the stop-line are estimated on the basis of the on-line detector data. Moreover, the current state of traffic, such as how many vehicles have been detected and how many of them have passed the stop-line in each stream can be estimated continuously over time. The model should represent the individual vehicle motion in relation to the general car-following concept. Thus, it can be applied in the dynamic signal optimisation at a microscopic level.

4.2 KINEMATIC EQUATIONS IN ONE-DIMENSION

Kinematics is the study of motion irrespective of the forces; it deals with the mathematical description of motion in terms of position, speed, acceleration and time. If any three of those variables are known, then the fourth variable can be calculated by using kinematic equations. Three types of motion are recognised in physics: translational, rotational and vibrational. A car moving along a street is an example of translational motion; the Earth's spin on its axis is an example of rotational motion; and the back-and-forth movement of a pendulum is an example of vibrational motion. Amongst these, we focus on the translational motion in one-dimension to describe the motion of vehicles in the vicinity of signalised junctions.

If a vehicle is moving, the speed v is defined as the displacement of the vehicle divided by the time over which the displacement occurs. Furthermore, acceleration a refers to the rate of change of speed over time, which is defined as the change of speed divided by the change of time. The acceleration is equal to the second derivative of x with respect to time t. The position, speed and acceleration as a function of time can be expressed as follows:

$x(t)$ is position as a function of time [m]

$v(t) = \frac{dx}{dt}$ is speed as a function of time [m/s]

$a(t) = \frac{dv}{dt} = \frac{d^2x}{dt^2}$ is acceleration as a function of time [m/s^2]

The simplest case of translational motion assumes that the acceleration is constant, which means all objects are moving with maximum uniform acceleration rate, where

$$a = \frac{dv}{dt} \quad (a \text{ is constant}) \tag{4.1}$$

Integrating (4.1) over the limits from initial time t_i to final time t_f and initial speed v_i to final speed v_f gives

$$\int_{v_i}^{v_f} dv = \int_{t_i}^{t_f} a\,dt \tag{4.2}$$

then, the speed equation as a function of time is given by

$$v_f = v_i + a(t_f - t_i) \tag{4.3}$$

From the description of an acceleration (4.1), applying the chain rule gives

$$\begin{aligned} a &= \frac{dv}{dx}\left(\frac{dx}{dt}\right) \\ &= \frac{dv}{dx}v \end{aligned} \tag{4.4}$$

Integrating (4.4) over the limits from initial speed v_i to final speed v_f and initial position x_i and x_f gives

$$\int_{v_i}^{v_f} v\,dv = \int_{x_i}^{x_f} a\,dx \tag{4.5}$$

then, the position equation as a function of speed is given by

$$x_f = x_i + \frac{v_f^2 - v_i^2}{2a} \tag{4.6}$$

Using (4.3) in (4.6), we can get the position equation as a function of time:

$$x_f = x_i + v_i(t_f - t_i) + \frac{a}{2}(t_f - t_i)^2 \tag{4.7}$$

The kinematic equations are a set of four different variables: position, speed, acceleration (or braking) and time, if the initial conditions at time t_i are denoted, respectively by position X_i and speed v_i, the final position X_f and speed v_f at time t_f are calculated with constant acceleration rate a or braking rate b. In this manner, the motion of vehicle is completely known if the vehicle's position in space is known at all times. Using this concept, the motion of vehicles at signal controlled junctions from the position of detector to stop-line can be analytically formulated as a function of signal indications.

4.3 TRAJECTORY OF THE LEADING VEHICLE

In this section, a single vehicle is considered approaching a traffic signal along an empty road: this provides a model for the motion of a leading vehicle. Based on the general kinematic equations (4.3), (4.6) and (4.7), a kinematic car-following model for the leading vehicle is formulated analytically as a function of the start of green time.

4.3.1 Notation for the leading vehicle trajectory and calculation summary

1) Notation for the leading vehicle

The following notations will be used when considering a trajectory of the leading vehicle at signalised junctions between the detector and the stop-line with respect to the current signal indication; more additional notations will be required later for the motion of following vehicles.

Let

v_0 be the free-flow speed that the driver of the vehicle wishes to undertake (desired speed),

a be the acceleration rate that the vehicle undertakes ($a>0$),

b be the braking rate that the vehicle undertakes ($b>0$),

X_d be the position of the detector ($X_d<0$),

t_d be the time at which the vehicle is detected at position X_d,

v_d be the speed of the vehicle at time that it is detected at position X_d (suppose that $v_d=v_0$),

X_s be the position of the stop-line ($X_s=0$),

t_s be the time at which the vehicle crosses the stop-line X_s,

v_s be the speed of the vehicle at time that it crosses the stop-line X_s,

t_g be the time at which the vehicle makes its motion decision. When a signal display changes from the red and amber to the green, the time t_g is defined as a start of red and amber time plus a reaction time, which is the start of green time in KCS traffic model (not an indicated green time),

v_g be the speed of the vehicle at time t_g,

X_g be the position of the vehicle at time t_g ($X_g \le 0$),

X_b be the position at which the vehicle has to start braking in order to stop safely at position X_s, in the case that the current signal is red ($X_b<0$),

t_b be the time at which the vehicle starts to brake,

v_b be the speed of the vehicle at time that it reaches the braking position X_b (suppose that $v_b=v_0$),

X_q be the position at which the vehicle stops after a full braking ($X_q=X_s=0$),

t_q be the time at which the vehicle stops after a full braking (where the speed $v_q=0$),

X_v be the position at which any delayed vehicle regains the free-flow speed v_0,

t_v be the time at which any delayed vehicle regains the free-flow speed (where the speed $v_v=v_0$).

2) Calculation summary for the leading vehicle trajectory

The motion of the leading vehicle at signalised junction is subject to the current signal indication (i.e. green or red). In respect of the current signal indication, the motion can be classified broadly into four regions which are traversed in the same order from the detector through the stop-line: free-flow,

braking (or braking and stopping), acceleration and free-flow. The important feature of this motion is that acceleration always follows braking or stopping, which means that in the absence of braking or stopping free-flow speed is maintained.

We consider link a single vehicle approaching a traffic signal along an empty road. If there is no vehicle in front and current signal is green, under such circumstances the first vehicle from the detector will cross the detector with free-flow speed and then reaches stop-line without experiencing any delays. However, if the signal is currently red, the first vehicle from the detector can travel up to the braking position with free-flow speed, and on reaching that point the vehicle has to start to brake in order to stop safely at the stop-line. Meanwhile, if the signal changes to green, the vehicle will start to accelerate until it regains the free-flow speed; otherwise, the vehicle has to go through braking and stopping until the next green starts. In this way, we can identify that the vehicle will pass the stop-line either at the free-flow speed or under.

As seen in Figure 4.1, there are three different trajectories to be considered according to the relation of the time t_g with t_b and t_q. The three different trajectories are considered as follows:

- If $t_g \leq t_b$: maintain free-flow speed (see Figure 4.1a)
- If $t_b < t_g \leq t_q$: free-flow → braking → acceleration → free-flow (see Figure 4.1b)
- If $t_g > t_q$: free-flow → braking → stopping → acceleration → free-flow(see Figure 4.1c)

Here, the time t_g is the start of green time that will be used in the KCS traffic model. It is defined as the beginning of the red and amber time plus a reaction time τ (for example, Gipps used τ = 2/3 seconds). At signalised junctions, when the leading vehicle is moving towards the stop-line, the driver will make a decision (either brake, accelerate or maintain the current speed) according to the current signal display. For example, if the signal turns from the red to red and amber whilst the vehicle is braking, the driver will know that green is following soon and may react accordingly. From this time, the driver will start to accelerate after the reaction time, but there is no change in motion during the reaction time period. In this manner, the start of green time t_g uses in the KCS traffic model is adjusted from the beginning of red and amber. For countries that have no starting amber, it can be adjusted from the beginning of indicated green.

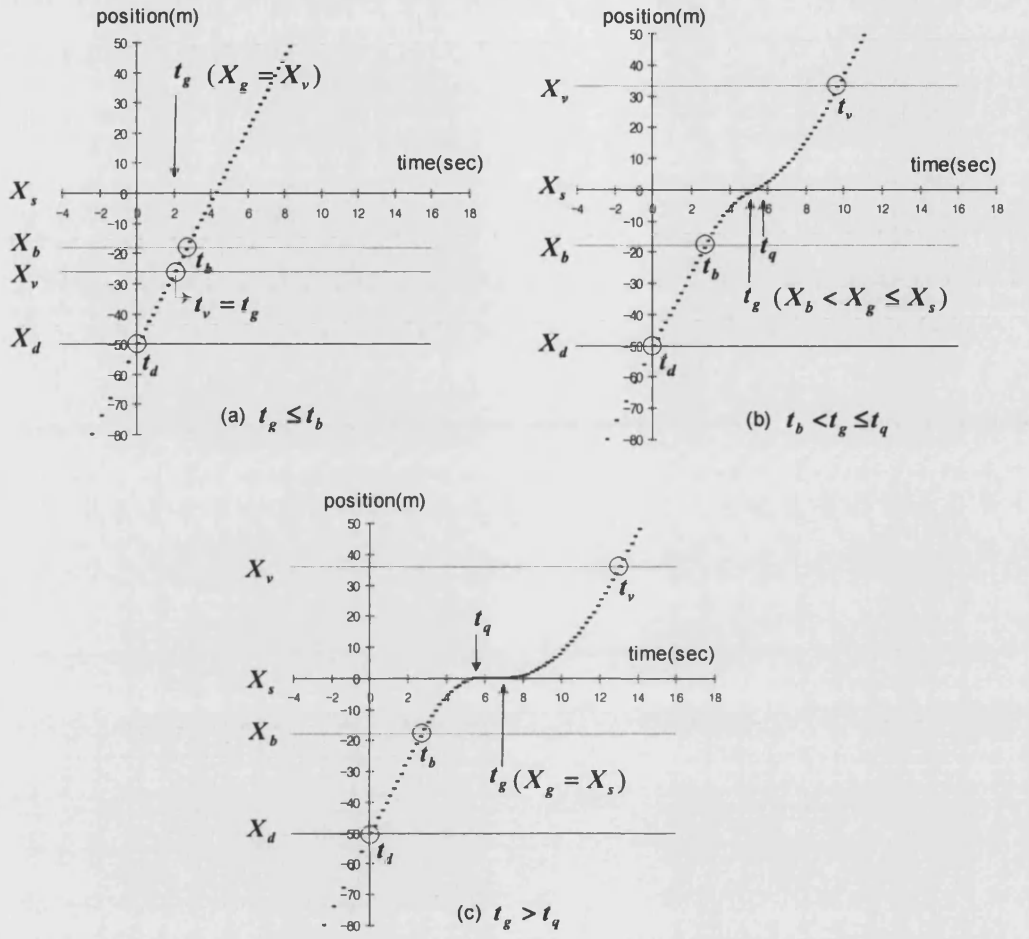


Figure 4.1 Trajectory of the single vehicle in relation to varying start of green t_g

In the following sections, the motion of leading vehicle is numerically formulated as a function of t_g with respect to the flow provided from the detector t_d. This model assumes a constant acceleration rate a and braking rate b ($a, b > 0$), no overtaking is allowed and no vehicle exceeds the free-flow speed v_0 under any circumstances. In addition, the position of the stop-line X_s ($X_s = 0$) and the detector X_d ($X_d < 0$) are the only fixed special parameters whilst the other quantities vary according to the signal displays. The present analysis can be applied directly to a single detector at any position upstream of the stop-line. Here, the position of detector is selected in accordance with the maximum number of vehicles that can be accommodated during a green period (e.g. maximum green period). If the detector were any closer to the stop-line, the present method could still be applied, but less information would be available on arrivals. Several detectors could be used on each approach, but

this would require further rules to update vehicle trajectories in the KCS traffic model. There are three different position variables (X_g, X_b and X_v) and two different time variables (t_b and t_v) that we need to identify from the detector through the stop-line by introducing an additional time parameter t_g.

4.3.2 The motion up to the braking position ($X_d \rightarrow X_b$)

As seen in Figure 4.1, the motion of the leading vehicle from the detector position up to the braking position is constant with free-flow speed: the motion within this region is unaffected by the current signal indication. In formulating a kinematic car-following model at signalised junctions, the first task is to define a braking position X_b and its time t_b, where the speed $v_b = v_0$, thus the vehicle comes to a halt at the stop-line if the signal changes to red.

1) The braking position X_b and its time t_b

The braking position X_b is given by

$$X_b = X_s - \frac{v_0^2}{2b} = -\frac{v_0^2}{2b} \tag{4.8}$$

and its time t_b is

$$t_b = t_d + \frac{(X_b - X_d)}{v_0} \tag{4.9}$$

using (4.8) in (4.9), we get

$$t_b = t_d - \frac{v_0}{2b} - \frac{X_d}{v_0} \tag{4.10}$$

2) The duration of full braking Δt_b and the stopping time t_q

Additional information that can be estimated from the position X_b is the duration of full braking Δt_b until the vehicle stops and its stopping time t_q at the stop-line. The duration of full braking Δt_b is given by

$$\Delta t_b = \frac{v_0}{b} \tag{4.11}$$

and the stopping time t_q is

$$t_q = t_b + \Delta t_b \tag{4.12}$$

using (4.10) and (4.11) in (4.12), the stopping time t_q becomes

$$t_q = t_d + \frac{v_0}{2b} - \frac{X_d}{v_0} \tag{4.13}$$

4.3.3 The motion after the braking position ($X_b \rightarrow X_v$)

When the vehicle has reached its braking position, its further motion is determined by current signal display at the junction. On reaching the braking position, the driver has to make a decision in respect of signal indication: if the signal is green the driver can maintain the free-flow speed as at the detector; however, if the signal is red then the driver has to start braking. Meanwhile, if the green starts during the braking, then the driver switches to acceleration until the vehicle regains the free-flow speed; otherwise, the vehicle has to go through full braking and then stop.

In Figure 4.1, there are three different trajectories to be considered with respect to the start of green time t_g; trajectory (a) shows the case of $t_g \leq t_b$, in which the vehicle is undelayed (can maintain free-flow motion), trajectory (b) shows the case of $t_b < t_g \leq t_q$, in which the vehicle goes through braking and then acceleration without stopping (free-flow → braking → acceleration → free-flow), and finally, trajectory (c) shows the case of $t_g > t_q$, in which the vehicle brakes, stops and then accelerates until it regains the free-flow speed (free-flow → braking → stopping → acceleration → free-flow). From now on following equations can be expressed as a function of t_g.

1) Speed v_g and position X_g at time t_g

- If $t_g \leq t_b$, we can suppose that the vehicle has the free-flow speed at time t_g, then

$$v_g = v_0 \quad \text{and} \quad X_g = X_d + v_0(t_g - t_d) \tag{4.14}$$

- If $t_b < t_g \leq t_q$, the motion of vehicle at time t_g is during the braking, then

$$v_g = v_0 - b(t_g - t_b) \tag{4.15}$$

and

$$X_g = X_b + v_0(t_g - t_b) - \frac{b}{2}(t_g - t_b)^2 \tag{4.16}$$

- If t_g starts after the vehicle has stopped ($t_g > t_q$), we can simply assume that

$$v_g = 0 \quad \text{and} \quad X_g = X_s = 0 \tag{4.17}$$

2) Time t_v and position X_v of regaining the free-flow speed

- If $t_g \le t_b$, the motion of vehicle is unaffected by t_g, thus we can simply assume that

$$t_v = t_g \tag{4.18}$$

and

$$X_v = X_g \tag{4.19}$$

- If $t_b < t_g \le t_q$, the braking vehicle will start to accelerate at time t_g, thus

$$t_v = t_g + \frac{v_0 - v_g}{a} \tag{4.20}$$

using (4.15) in (4.20), we get

$$t_v = t_g + \frac{b}{a}(t_g - t_b) \tag{4.21}$$

and

$$X_v = X_g + v_g(t_v - t_g) + \frac{a}{2}(t_v - t_g)^2 \tag{4.22}$$

using (4.15), (4.16) and (4.21) in (4.22), we get

$$X_v = X_b + v_0(\frac{a+b}{a})(t_g - t_b) - (\frac{ab+b^2}{2a})(t_g - t_b)^2 \tag{4.23}$$

- If t_g starts after the vehicle has stopped ($v_g = 0$), then

$$t_v = t_g + \frac{v_0}{a} \tag{4.24}$$

and

$$X_v = X_s + \frac{a}{2}(t_v - t_g)^2 = \frac{a}{2}(t_v - t_g)^2 = \frac{v_0^2}{2a} \tag{4.25}$$

Once we have derived numerical equations for the leading vehicle from the X_d to the X_v, then the crossing time and its speed to the stop-line can be obtained by comparing the fixed position X_s and the varying position X_v: whether X_v is before the stop-line or after the stop-line.

3) Crossing time t_s and its speed v_s to the stop-line $X_s = 0$

- If $X_v \le X_s$, which means that the vehicle is crossing the stop-line with free-flow speed, then

$$t_s = t_v + \frac{X_s - X_v}{v_0} = t_v - \frac{X_v}{v_0} \quad \text{and} \quad v_s = v_0 \tag{4.26}$$

- If $X_v > X_s$, which means that the vehicle is crossing the stop-line during the acceleration, then the speed v_s can be calculated according to the equation of

$$X_s = X_g + \frac{v_s^2 - v_g^2}{2a} \tag{4.27}$$

from (4.27), the speed v_s can be obtained by

$$v_s = \sqrt{v_g^2 - 2aX_g} \tag{4.28}$$

and then we can get

$$t_s = t_g + \frac{v_s - v_g}{a} = t_g + \frac{\sqrt{v_g^2 - 2aX_g} - v_g}{a} \tag{4.29}$$

4.3.4 Discussion

Here, we have established KCS traffic model for the leading vehicle. All variables (time, position and speed) are estimated based on the detection time with respect to the start of the green time. The resulting estimates will be used as parameters of the leading vehicle in the following Section 4.4 (Trajectory of following vehicles).Then, afterwards motion for the following vehicles can be determined based on the leading vehicle trajectory with respect to the information provided from the detector and their possible departure time at downstream free-flow position.

In the UK, the cyclic order of signal displays is red → red and amber → green → amber. Thus, the green follows after the 2 seconds of the red and amber period. In such case, the start of green time t_g including reaction time is adjusted from the start of red and amber. However, some countries do not have a red and amber period (starting amber), then the start of green time t_g should be adjusted from the beginning of indicated green time to use in the KCS traffic model.

4.4 TRAJECTORY OF THE FOLLOWING VEHICLES

In this section, KCS traffic model for the following vehicles is developed on the basis of variables estimated from the leading vehicle trajectory. Once we have characterised the full trajectory of the leading vehicle as a function of the start of green time, the trajectory of all successive following vehicles can be calculated directly using the information provided from the detector. The general and basic concepts we use in the following calculations are: the motion of the following vehicles are determined by the possible departure time at downstream free-flow position and they cannot depart this position with less than the minimum headways.

4.4.1 The additional notation and calculation summary

1) Notation for the following vehicle trajectory

The notation will be used for the trajectory of the following vehicle $n\,(2 \le n \le N)$, where N is the serial number of the most recent detected vehicle. Notations of the leading vehicle described in Section 4.3.1 will be considered as parameters in the following section.

Let

t_d^n be the time at which following vehicle n is detected at position X_d,

v_d^n be the speed of following vehicle n at time which it detected at position X_d,

t_s^n be the time at which following vehicle n crosses the stop-line X_s,

v_s^n be the speed of following vehicle n at time which it crosses the stop-line X_s,

t_a^n be the time at which following vehicle n starts to accelerate,

v_a^n be the speed of following vehicle n at time which the acceleration starts,

X_a^n be the position of following vehicle n when it starts to accelerate ($X_a^n \le 0$),

X_b^n be the braking position of following vehicle n, if sudden braking is necessary for a vehicle n-1 in front, from that position the following vehicle can stop safely at position X_q^n ($X_b^n = X_b^{n-1} - L$, where L is the safety margin),

t_b^n be the time at which following vehicle n starts to brake,

v_b^n be the speed of following vehicle n at the time when it reaches the braking position X_b^n ($v_b^n = v_0$),

X_q^n be the position at which following vehicle n stops completely after braking fully,

t_q^n be the time at which following vehicle n stops,

v_q^n be the speed of following vehicle n at the time when it reaches the stopping position X_q^n ($v_q^n = 0$),

X_v^n be the position at which following vehicle n regains the free-flow speed, if it has been delayed $X_a^n < X_v^n \le \overline{X}_v$; otherwise, all position variables are identical at $X_b^n = X_a^n = X_v^n$

t_v^n be the time at which following vehicle n regains the free-flow speed, if it has been delayed $t_a^n = t_v^n = \overline{t}_v^n$; otherwise, all time variables are identical at $t_b^n = t_a^n = t_v^n$,

$\overline{X}_v$ be the maximum boundary of downstream at which any delayed or stopped vehicle can regain the free-flow speed. This position is determined by acceleration rate a (a>o) and free-flow speed v_0; the possible departure time $\overline{t}_v^n$ for the following vehicle n will be calculated at this position ($\overline{X}_v = X_s + v_0^2 / 2a$),

$\overline{t}_{d,v}^n$ be the free-flow travel time of the following vehicle n from the detector to the $\overline{X}_v$

$(\bar{t}^n_{d,v} = t^n_d + (\bar{X}_v - X_d)/v_0)$,

$\hat{t}^n_v$ be the earliest departure time of following vehicle n; previous vehicle's departure time plus the minimum headway $1/s$ $(\hat{t}^n_v = \bar{t}^{n-1}_v + 1/s)$,

$\bar{t}^n_v$ be the possible departure time of following vehicle n at position $\bar{X}_v$ $(\bar{t}^n_v = Max[\hat{t}^n_v, \bar{t}^n_{d,v}])$,

$\bar{t}^n_{d,s}$ be the free-flow travel time of the following vehicle n from the detector to the stop-line X_s $(\bar{t}^n_{d,s} = t^n_d + (X_s - X_d)/v_0)$,

$\hat{t}^n_{b,v}$ be the longest approach time from the braking position X^n_b to the free-flow position $\bar{X}_v$ supposing that the vehicle does not stop, which is called a standard motion arrival time.

$\bar{X}_b$ be the maximum boundary of upstream; if any braking position X^n_b is identified beyond that position, the motion only can be described from the detector, because it is not possible to estimate what happened in the past, likewise, too far to estimate the motion from the detector $(\bar{X}_b = X_d - v_0^2/2b)$.

2) Calculation summary for the following vehicle trajectory

The variables estimated for the leading vehicle in the Section 4.3 are used as parameters for the following vehicles trajectory. Once we have characterised the full trajectory of the leading vehicle n=1 as a function of the start of green time t_g, the trajectory of all successive following vehicles n=2,3,... N can be calculated directly based on the information provided from the detector. The basic concept we use in the following calculation is that any successive following vehicle cannot depart the downstream position $\bar{X}_v$ with less than minimum headways.

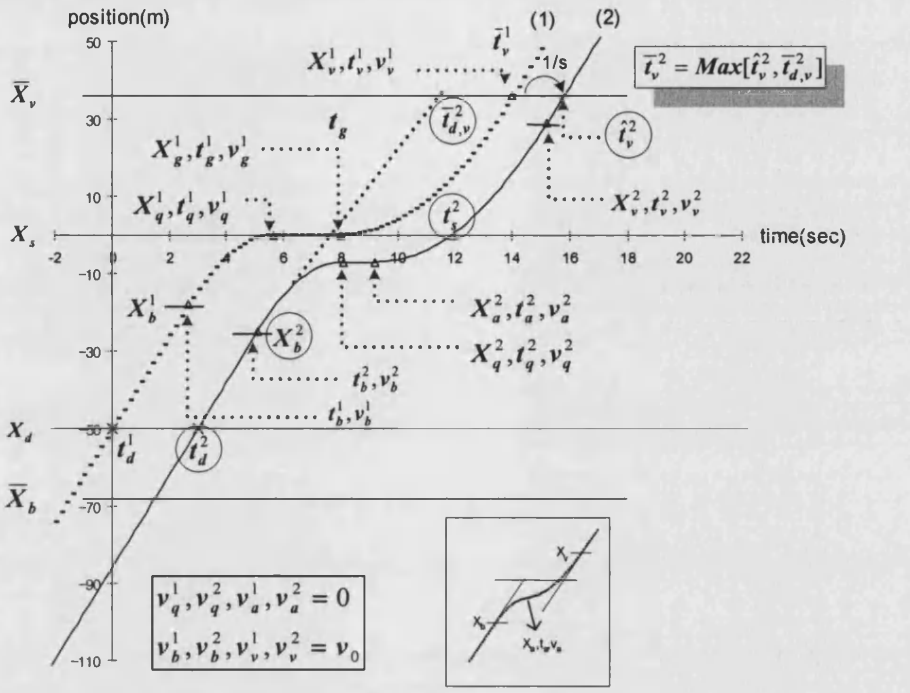


Figure 4.2 Position, time and speed variables for the following vehicle *n*=2

As seen in Figure 4.2, when the following vehicle $n=2$ crosses the detector X_d at time t_d^2, the first order of task is finding a possible departure time $\overline{t}_v^2$ and its braking position X_b^2. At position $\overline{X}_v$, by comparing variables of the earliest departure time $\hat{t}_v^2$ and the free-flow arrival time $\overline{t}_{d,v}^2$, the possible departure time will be $\overline{t}_v^2 = \text{Max}[\hat{t}_v^2, \overline{t}_{d,v}^2]$, in which we can decide whether or not the following vehicle will be delayed. The braking position X_b^2 is calculated by of adding minimum safe spacing L, each of which is calculated according to the trajectory of the leading vehicle.

If the detected vehicle is identified as undelayed, it can reach the downstream boundary $\overline{X}_v$ from the detector with free-flow arrival time. Thus, it is not necessary to find acceleration variables. However, if the vehicle is identified as delayed, then we need standard motion test (see Section 4.4.5.1) to find acceleration variables, such as t_a^2, X_a^2 and v_a^2. In order to find acceleration variables, we need to test whether or not any stopped lost time due to a queue is involved. Finally, we can calculate other quantities, such as $t_v^2, X_v^2, t_s^2, v_s^2$ and v_d^2.

In this analysis, the final and most important information we are seeking for each following vehicle $n(2 \le n \le N)$ is its crossing time t_s^n at the stop-line X_s and departure time $\overline{t}_v^n$ at the down stream free-flow position $\overline{X}_v$: the time t_s^n will be used to estimate the number of vehicles that can pass the stop-line if the green is extended by a certain control decision time, and $\overline{t}_v^n$ will be used to estimate its delays.

The calculation summary for the following vehicle trajectory is as follows ($2 \le n \le N$):

Step 0: (Leading vehicle trajectory)

Identify the leading vehicle trajectory n=1 with respect to current signal indications. All variables identified in this step are used as parameters in the following steps.

$\Rightarrow$ if N=1, stop processing (no following vehicles).

Step 1: (Departure time at $\overline{X}_v$)

Find the earliest departure time $\hat{t}_v^n$ and free-flow arrival time $\overline{t}_{d,v}^n$ at position $\overline{X}_v$, then the possible departure time will be $\overline{t}_v^n = \text{Max}[\hat{t}_v^n, \overline{t}_{d,v}^n]$,

where, $\hat{t}_v^n = \overline{t}_v^{n-1} + 1/s$ and $\overline{t}_{d,v}^n = t_d^n + (\overline{X}_v - X_d)/v_0$.

Step 2: (Motion definition for undelayed case or delayed case)

Find the expected motion of the each following vehicle, whether or not it will be delayed:

2-1) If $\overline{t}_{d,v}^n \ge \hat{t}_v^n$ which corresponds to undelayed motion, then $\overline{t}_v^n = \overline{t}_{d,v}^n$.

2-2) If $\overline{t}_{d,v}^n < \hat{t}_v^n$ which corresponds to delayed motion, then $\overline{t}_v^n = \hat{t}_v^n$,

a standard motion test is required to identify that the motion will be a halt case or not-halt case.

Step 3: (Braking position and time)

3-1) The braking position X_b^n with respect to minimum safe spacing for each following vehicle is calculated by $X_b^n = X_b^{n-1} - L$.

3-2) The braking time t_b^n is only possible to find if $X_b^n \ge \overline{X}_b$; otherwise, out of range.

- From Step 2, if the motion is identified as an undelayed case:

If $(X_b^n \geq \overline{X}_b)$, then $t_b^n = t_d^n + (X_b^n - X_d)/v_0$,

else, t_b^n is unknown.

$\Rightarrow$ go to Step 4 to find the final variable of crossing the stop-line.

- From Step 2, if the motion is identified as a delayed case:

If $X_d \leq X_b^n < X_s$, then $t_b^n = t_d^n + \dfrac{X_b^n - X_d}{v_0}$.

If $\overline{X}_b \leq X_b^n < X_d$, then $t_b^n = t_d^n - \left(\dfrac{v_0 - \sqrt{v_0^2 + 2b(X_b^n - X_d)}}{b} \right)$.

$\Rightarrow$ go to Step 5 to find additional variables.

Step 4: (Crossing time to the stop-line for the undelayed vehicle)

The crossing time t_s^n to the stop-line X_s can be directly calculated from the detector by using a free-flow motion equation, that is $t_s^n = t_d^n + (X_s - X_d)/v_0$.

Other position and time variables are identical to braking variables: $X_a^n = X_b^n, X_v^n = X_b^n$, $t_a^n = t_b^n$ and $t_v^n = t_b^n$.

$\Rightarrow$ if $n < N$, go to Step 2; otherwise, stop processing.

Step 5: (Standard motion test for the delayed vehicle: halt case or not-hale case)

The standard motion test is only necessary if the vehicle is identified as delayed and $X_b^n \geq \overline{X}_b$. From the braking position with known time t_b^n, before the vehicle starts to brake, standard motion test is needed to find whether or not it will come to a halt.

5-1) The standard motion arrival time $\hat{t}_{b,v}^n = t_b^n + v_0(a+b)/2ab + (\overline{X}_v - X_b^n)/v_0$, which is the longest approaching time from X_b^n to $\overline{X}_v$, supposing that the vehicle accelerates at time $t_a^n = t_q^n$.

5-2) Motion definition: halt-case or not-halt case

- If $\hat{t}_{b,v}^n < \overline{t}_v^n$, the motion is identified as a halt-case (see Figure 4.6 a), then the motion will be: braking → stopping → acceleration → free-flow.

- If $\hat{t}_{b,v}^n \geq \bar{t}_v^n$, the motion is identified as a not-halt case (see Figure 4.6 b), then the motion will be: braking → acceleration → free-flow without stopping.

Step 6: (Acceleration position, time and speed variables for the delayed vehicle)

6-1) For a halt-case, acceleration position is equal to the stopped position, then

$$X_a^n = X_q^n \text{ (where, } X_q^n = X_b^n + \frac{v_0^2}{2b}\text{), } v_a^n\text{=0 and } t_a^n = \bar{t}_v^n - \frac{v_0}{2a} - \frac{(\bar{X}_v - X_a^n)}{v_0}$$

6-2) For a not-halt case, acceleration starts between braking and stopped position, then

$$t_a^n = t_b^n + \sqrt{\frac{(X_b^n - \bar{X}_v) + v_0(\bar{t}_v^n - t_b^n)}{(ab+b^2)/2a}},\ X_a^n = X_b^n + v_0(t_a^n - t_b^n) - \frac{b}{2}(t_a^n - t_b^n)^2$$

and

$$v_a^n = v_0 - b(t_a^n - t_b^n).$$

Step 7: (Position and time of regaining the free-flow speed for the delayed vehicle)

If $X_b^n \geq \bar{X}_b$, the time t_v^n and the position X_v^n of regaining the free-flow speed can be calculated by using known acceleration variables X_a^n, t_a^n and v_a^n; otherwise, they can be calculated corresponding to the possible departure time $\bar{t}_v^n$ at $\bar{X}_v$.

7-1) If $(X_b^n \geq \bar{X}_b)$

$$t_v^n = t_a^n + \frac{v_0 - v_a^n}{a} \text{ and } X_v^n = X_a^n + v_a^n(t_v^n - t_a^n) + \frac{a}{2}(t_v^n - t_a^n)^2.$$

7-2) If $(X_b^n < \bar{X}_b)$ which is defined as a out of boundary case, then

$$t_v^n = t_d^n + \sqrt{\frac{2\left[(X_d - \bar{X}_v) + v_0(\bar{t}_v^n - t_d^n)\right]}{a}} \text{ and}$$

$$X_v^n = X_d + v_d^n(t_v^n - t_d^n) + \frac{a}{2}(t_v^n - t_d^n)^2$$

where the estimated speed of $v_d^n = v_0 - a(t_v^n - t_d^n)$.

Step 8: (Crossing time at the stop-line for the delayed vehicle)

This is the final information we are seeking for from the following vehicles trajectory. The arrival time t_s^n at the stop-line X_s can be calculated by comparing position variables X_s and X_v^n:

8-1) If $(X_b^n \geq \bar{X}_b)$

If $X_v^n \leq X_s$, then $t_s^n = t_v^n - \frac{X_v^n}{v_0}$.

If $X_v^n > X_s$, then $t_s^n = t_a^n + \frac{v_s^n - v_a^n}{a}$, where $v_s^n = \sqrt{(v_a^n)^2 - 2aX_a^n}$.

8-2) If $(X_b^n < \bar{X}_b)$

If $X_v^n \leq X_s$, then $t_s^n = t_v^n - \frac{X_v^n}{v_0}$.

If $X_v^n > X_s$, then $t_s^n = t_v^n + \frac{v_0 - v_s^n}{a}$, where $v_s^n = \sqrt{(v_d^n)^2 - 2aX_d}$.

$\Rightarrow$ if $n < N$, go to Step 2; otherwise, stop processing.

4.4.2 Departure time at position $\bar{X}_v$ and the motion definition

The position $\bar{X}_v$ is the maximum boundary of downstream at which any delayed or stopped vehicles will regain the free-flow speed, that position is determined by acceleration rate a and free-flow speed v_0:

$$\bar{X}_v = X_s + \frac{v_0^2}{2a} = \frac{v_0^2}{2a} \tag{4.30}$$

Each vehicle has different free-flow position. If the leading vehicle was waiting in the queue, it will regain the free-flow speed at position $X_v^1 = \bar{X}_v$ after a full acceleration; otherwise, always $X_v^1 \leq \bar{X}_v$. In the interest of simplicity, the motion of all vehicles are extended until the position $\bar{X}_v$, at which any undelayed or delayed vehicles will have regained the free-flow speed. From equations (4.18), (4.20) and (4.24), the extended arrival time $\bar{t}_v^1$ at the position $\bar{X}_v$ can be obtained by adding an additional free-flow travel time from X_v^1 (equations 4.19, 4.22 and 4.25) to $\bar{X}_v$. Thus, the arrival time of the leading vehicle $n = 1$ at $\bar{X}_v$ becomes $\bar{t}_v^n = t_v^n + (\bar{X}_v - X_v^n)/v_0$.

At the position $\bar{X}_v$, the basic assumption we need to apply is that any successive vehicles cannot depart this position with less than the minimum saturation departure time $1/s$ .Then, the earliest departure time $\hat{t}_v^n$ at $\bar{X}_v$ is estimated by

$$\hat{t}_v^n = \bar{t}_v^{n-1} + 1/s \qquad (2 \le n \le N) \tag{4.31}$$

The free-flow travel time $\bar{t}_{d,v}^n$ from the detector to the $\bar{X}_v$ is given by

$$\bar{t}_{d,v}^n = t_d^n + \frac{(\bar{X}_v - X_d)}{v_0} \qquad (2 \le n \le N) \tag{4.32}$$

By comparing the time variables obtained from equations (4.31) and (4.32), the possible departure time $\bar{t}_v^n$ at $\bar{X}_v$ can be determined by

$$\bar{t}_v^n = \text{Max}[\hat{t}_v^n, \bar{t}_{d,v}^n] \qquad (2 \le n \le N) \tag{4.33}$$

Now, we can identify that the following vehicle n ($2 \le n \le N$) is whether or not delayed by comparing two time variables of $\bar{t}_{d,v}^n$ and $\hat{t}_v^n$:

- If $\bar{t}_{d,v}^n \ge \hat{t}_v^n$ which corresponds to the undelayed motion, then $\bar{t}_v^n = \bar{t}_{d,v}^n$: we can suppose that the vehicle will crosses the stop-line with free-flow trajectory, thus, in that case seeking acceleration variables are not necessary.
- However, if $\bar{t}_{d,v}^n < \hat{t}_v^n$ which corresponds to the delayed motion, then $\bar{t}_v^n = \hat{t}_v^n$: we need standard motion test to find acceleration variables.

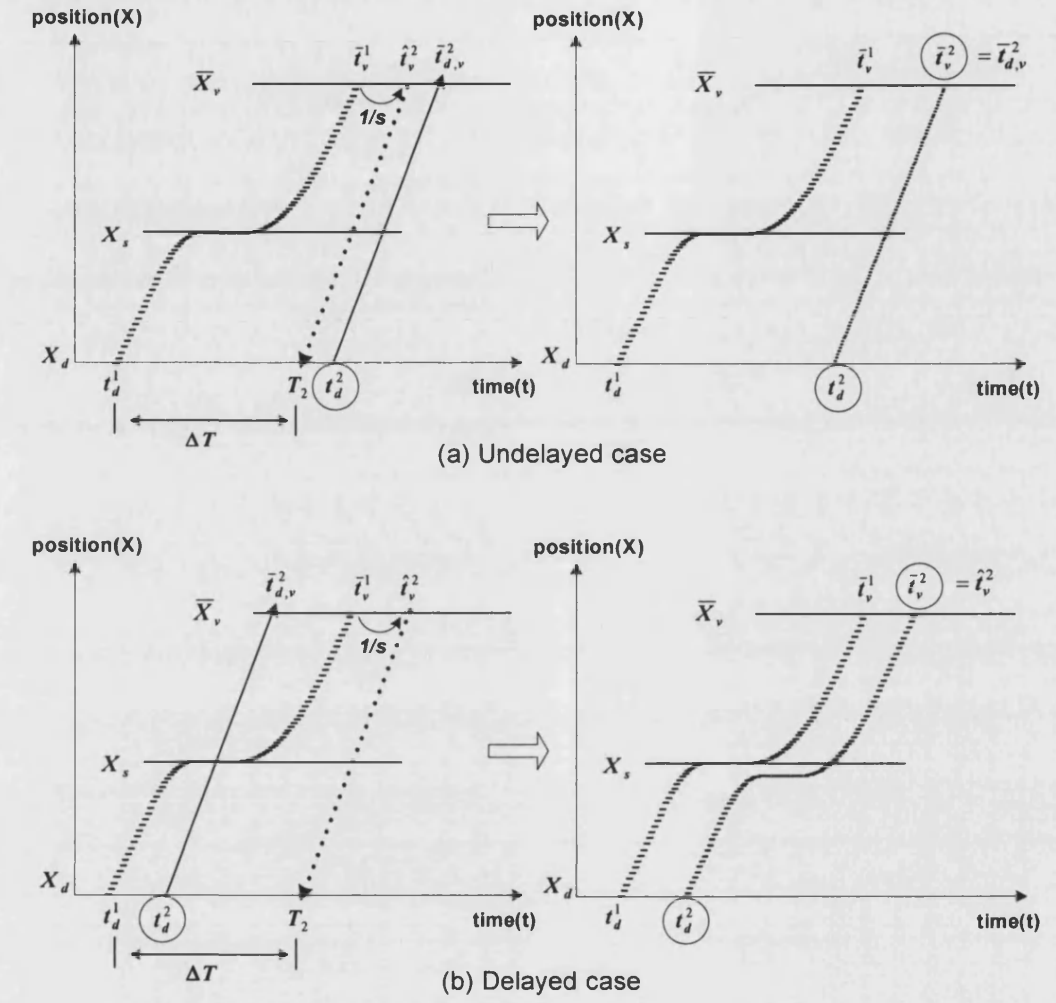


Figure 4.3 Departure time at the downstream position $\overline{X}_v$

As seen in Figure 4.3, when the following vehicle n=2 crosses the detector at time t_d^2, firstly we need to find the earliest departure time $\hat{t}_v^2$ and the free-flow arrival time $\overline{t}_{d,v}^2$, where $\hat{t}_v^2$ is calculated based on the previous vehicle's departure time (4.31), and $\overline{t}_{d,v}^2$ is directly estimated from the detector position (4.32). If the vehicle is detected after $t_d^1+\Delta T$, it would pass $\overline{X}_v$ some time after $\hat{t}_v^2$ and can leave the junction without experiencing any delay (see Figure 4.3a). However, if the vehicle is detected within a delay zone $\Delta T = T_2 - t_d^1$ (where $T_2 = \hat{t}_v^2 - (\overline{X}_v - X_d)/v_0$) its trajectory will eventually join the line segment $\overline{\hat{t}_v^2, T_2}$ and then will depart at time $\hat{t}_v^2$ with delay of $\hat{t}_v^2 - \overline{t}_{d,v}^2$ (see

Figure 4,3b). In this way, if $t_d^2 > T_2$, then $\overline{t}_v^2 = \overline{t}_{v,d}^2$ (undelayed case), and if $t_d^1 < t_d^2 \le T_2$, then $\overline{t}_v^2 = \hat{t}_v^2$ (delayed case) in accordance with the detector time.

4.4.3 Braking position and time

The braking position X_b^n for the following vehicle n ($2 \le n \le N$) is calculated by of adding a safety margin L on to the previous vehicle's braking position X_b^{n-1}. This will progress towards upstream detector until the end of the current stage. Each following vehicle has its own braking position, but the braking time t_b^n is only possible to find if $X_b^n \ge \overline{X}_b$. The position $\overline{X}_b$ is the maximum boundary of the upstream, that position is determined by the braking rate b and the free-flow speed v_0:

$$\overline{X}_b = X_d - v_0^2 / 2b \tag{4.34}$$

If any braking position occurs upstream of the position $\overline{X}_b$, the motion only can be described from the detector position towards to down stream, because the vehicle may had went through both braking and accelerating before reaching the detector (see Figure 4.5a, vehicle (9)).

The following vehicles braking position X_b^n are calculated by allowing a safety margin L, which includes the length of the vehicle:

$$X_b^n = X_b^{n-1} - L \quad (2 \le n \le N) \tag{4.35}$$

The braking time t_b^n can be calculated in two different ways (undelayed case and delayed case) by comparing X_b^n with three known position parameters, such as X_s, X_d and $\overline{X}_b$.

1) For the undelayed case

- If $\overline{t}_{d,v}^n \ge \hat{t}_v^n$, which corresponds to undelayed motion, then we can suppose that the vehicle is approaching the braking position through the detector with free-flow speed. The braking time t_b^n is calculated as follows:

If $\bar{X}_b \le X_b^n < X_s$, then

$$t_b^n = t_d^n + \frac{X_b^n - X_d}{v_0} \qquad (2 \le n \le N) \tag{4.36}$$

If the braking position is beyond the upstream boundary ($X_b^n < \bar{X}_b$), it is out of range, so that the time t_b^n is unknown.

2) For the delayed case

- If $\bar{t}_{d,v}^n < \hat{t}_v^n$, which corresponds to delayed motion, then the braking time can be calculated in two different ways: if X_b^n is between the stop-line and the detector, t_b^n is calculated same as (4.37), but if X_b^n is upstream of the detector which indicates that the vehicle has speed at the detector less than free-flow speed, so the vehicle is approaching the detector during the braking. The braking time t_b^n is calculated as follows:

If $X_d \le X_b^n < X_s$, then

$$t_b^n = t_d^n + \frac{X_b^n - X_d}{v_0} \qquad (2 \le n \le N) \tag{4.37}$$

If $\bar{X}_b \le X_b^n < X_d$, then

$$t_b^n = t_d^n - \left(\frac{v_0 - \sqrt{v_0^2 + 2b(X_b^n - X_d)}}{b} \right) \qquad (2 \le n \le N) \tag{4.38}$$

If the braking position is beyond the upstream boundary ($X_b^n < \bar{X}_b$), it is out of range, so that the time t_b^n is unknown.

4.4.4 Finding additional variables for the undelayed case

As discussed in Section 4.4.2, if the motion is defined as the undelayed case, then we can suppose that the vehicle will travel from the detector to the downstream position $\bar{X}_v$ with free-flow speed. In

that case, seeking acceleration variables for the following vehicle are not necessary, because acceleration always follows after braking or stopping, which means that in the absence of delay free-flow speed is maintained. Thus, other position and time variables for the following vehicle $n\,(2 \le n \le N)$ are all identical to the variables found in Section 4.4.3 (braking position and time for the undelayed case): $X_a^n = X_b^n$, $X_v^n = X_b^n$, $t_a^n = t_b^n$ and $t_v^n = t_b^n$.

4.4.4.1 Crossing time to the stop-line

The final information we are looking for from the following vehicle trajectory is its crossing time to the stop-line. When the motion has been defined as the undelayed case, this time variable can be calculated directly using the detector time t_d^n. The crossing time t_s^n to the stop-line X_s is given by

$$t_s^n = t_d^n + \frac{X_s - X_d}{v_0} = t_d^n - \frac{X_d}{v_0} \qquad (2 \le n \le N) \tag{4.39}$$

4.4.5 Finding additional variables for the delayed case

The following Sections 4.4.5.1 - 4.4.5.4 describe methods for finding the additional variables for a vehicle that has been identified as a delayed case: from the braking position towards the downstream boundary $\overline{X}_v$.

4.4.5.1 Standard motion test to identify halt case or not-halt case

The standard motion test is only necessary for the following vehicle n $(2 \le n \le N)$ that has been identified as having the delayed motion, and its braking position is within the upstream boundary $\overline{X}_b$ $(X_b^n \ge \overline{X}_b)$. This test will identify whether or not any stopped time is involved due to a queue is involved. It is an important test to find acceleration variables, and then other following variables can be calculated.

As seen in Figure 4.4, the standard arrival time $\hat{t}_{b,v}^2$ is the longest approach time from the braking position X_b^2 to the free-flow position $\overline{X}_v$, supposing that the vehicle does not stop $(t_a^2 = t_q^2)$. By comparing two time variables $\hat{t}_{b,v}^2$ and $\overline{t}_v^2$, the motion can be distinguished into one of two groups:

if any stopped delay is involved, the motion is called a halt case (see Figure 4.4a); otherwise, it is called a not-halt case (see Figure 4.4b). Hence, we can identify when, where and with which speed the following vehicle will start to accelerate, and additionally, the extra information of how long the vehicle had stayed in the queue until crossing the stop-line if the vehicle had to halt.

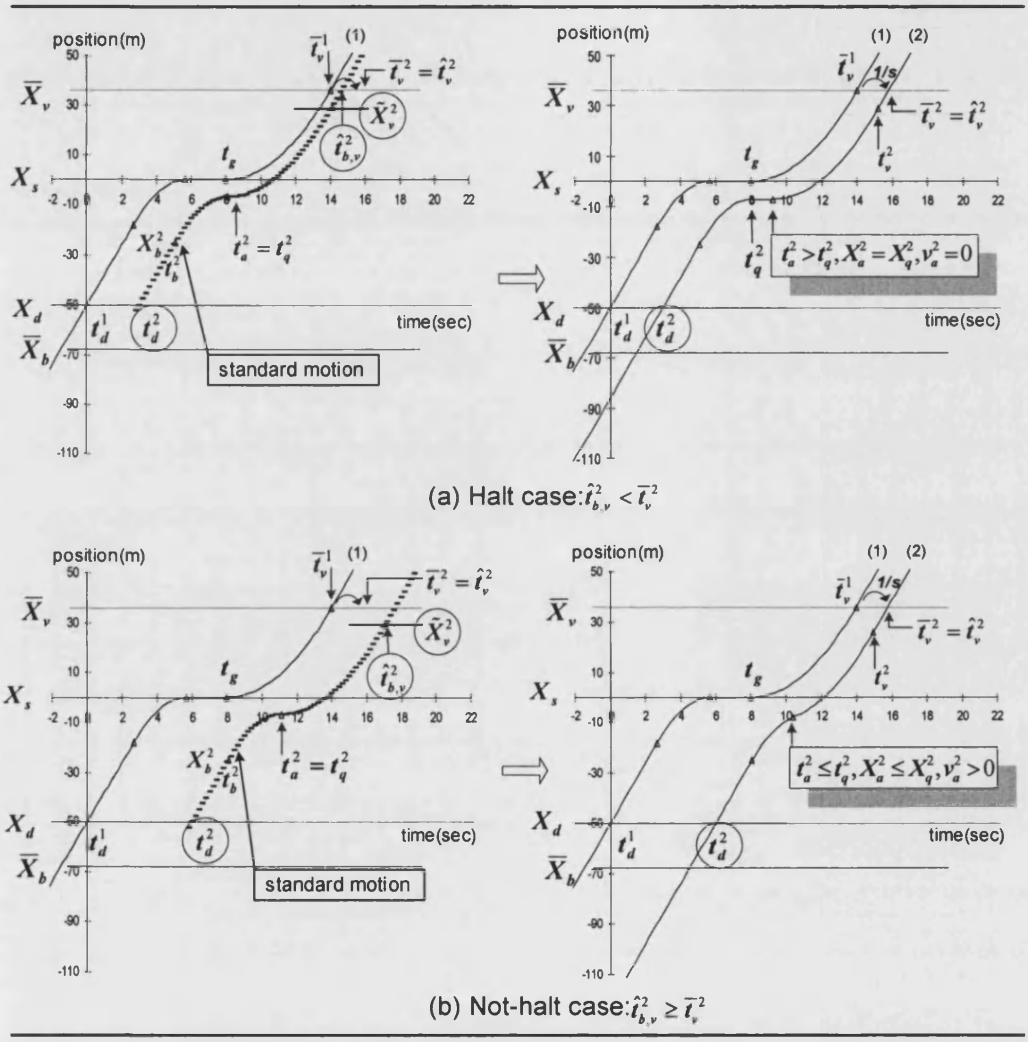


Figure 4.4 Standard motion test (halt case and not-halt case motion)

1) Standard motion arrival time

The standard motion arrival time for the following vehicle $n\,(2 \le n \le N)$ is calculated from the braking position X_b^n if the motion is identified as delayed case and t_b^n is known. The standard motion arrival time $\hat{t}_{b,v}^n$ at the position $\bar{X}_v$ is given by

$$\hat{t}_{b,v}^{n} = t_b^n + \Delta t_b + \Delta t_a + (\bar{X}_v - \tilde{X}_v^n)/v_0$$
$$= t_b^n + v_0(a+b)/ab + (\bar{X}_v - \tilde{X}_v^n)/v_0$$
$$= t_b^n + v_0(a+b)/2ab + (\bar{X}_v - X_b^n)/v_0 \qquad (4.40)$$

where

Δt_b is v_0/b, the duration of full braking from $v=v_0$ to $v=0$,

Δt_a is v_0/a, the duration of full acceleration from $v=0$ to $v=v_0$,

$\tilde{X}_v^n$ is the position after full braking and full acceleration ($\tilde{X}_v^n < \bar{X}_v$);

$$\tilde{X}_v^n = X_b^n + \frac{v_0^2}{2a} + \frac{v_0^2}{2b} = X_b^n + v_0^2\left(\frac{a+b}{2ab}\right).$$

2) Motion definition (halt case or not-halt case)

By comparing the standard arrival time $\hat{t}_{b,v}^n$ and the possible departure time $\bar{t}_v^n$, the motion of the following vehicle $n\,(2 \le n \le N)$ can be identified before calculating acceleration variables:

- If $\hat{t}_{b,v}^n < \bar{t}_v^n$, then this vehicle has to be held in its stopped position until the proper time for it to start acceleration can be obtained. This case is considered as a halt case, and hence $t_a^n > t_q^n$, $X_a^n = X_q^n$, and $v_a^n = 0$ (see figure 4.4 a).

- In contrast, if $\hat{t}_{b,v}^n \ge \bar{t}_v^n$, which means this vehicle does not have to be held in the stopped position, hence it will start to accelerate meanwhile of braking. In such case is considered as a not-halt case, and hence $t_a^n \le t_q^n, X_a^n \le X_q^n$ and $v_a^n > 0$ (see Figure 4.4b).

4.4.5.2 Acceleration time, position and speed

1) Finding acceleration variables for the halt case: (If $\hat{t}_{b,v}^n < \bar{t}_v^n$)

For the halt case, the acceleration position is equal to the stopping position, hence $X_a^n = X_q^n$ and the speed at this position is $v_a^n = 0$. The acceleration position X_a^n is given by

$$X_a^n = X_q^n = X_b^n + \frac{v_0^2}{2b} \qquad (2 \le n \le N) \tag{4.41}$$

From this position (4.42), the vehicle starts to accelerate at time t_a^n after a certain period of time lag, which is determined by the known departure time $\overline{t}_v^n$ at position $\overline{X}_v$. The acceleration time t_a^n can be calculated subject to the following three equations:

$$v_0 = a(t_v^n - t_a^n), \quad X_v^n = X_a^n + \frac{a}{2}(t_v^n - t_a^n)^2 \quad \text{and} \quad \overline{X}_v = X_v^n + v_0(\overline{t}_v^n - t_v^n)$$

then, the acceleration time t_a^n is given by

$$t_a^n = \overline{t}_v^n - \frac{v_0}{2a} - \frac{(\overline{X}_v - X_a^n)}{v_0} \qquad (2 \le n \le N) \tag{4.42}$$

The important information we can assume for the halt case is that some vehicles are still in halt, so the following vehicle will be held in the queue for a time $t_a^n - t_q^n (t_a^n > t_q^n)$.

2) Finding acceleration variables for the not-halt case: (If $\hat{t}_{b,v}^n \ge \overline{t}_v^n$)

The not-halt case calculation is more complicated than the halt case, because the vehicle start to accelerate immediately from braking, so that the acceleration speed is $v_a^n > 0$. In such a case, acceleration variables are calculated by considering simultaneously with given time, position and speed variables between the position X_b^n and $\overline{X}_v$. The acceleration time t_a^n can be calculated subject to the following five equations:

$$v_a^n = v_0 - b(t_a^n - t_b^n), \quad X_a^n = X_b^n + v_0(t_a^n - t_b^n) - \frac{b}{2}(t_a^n - t_b^n)^2,$$

$$v_0 = v_a^n + a(t_v^n - t_a^n), \quad X_v^n = X_a^n + v_0(t_v^n - t_a^n) + \frac{a}{2}(t_v^n - t_a^n)^2 \quad \text{and}$$

$$\overline{X}_v = X_v^n + v_0(\overline{t}_v^n - t_v^n)$$

then, the acceleration time t_a^n, position X_a^n and speed v_a^n are calculated as follows ($2 \le n \le k$):

$$t_a^n = t_b^n + \sqrt{\frac{(X_b^n - \bar{X}_v) + v_0(\bar{t}_v^n - t_b^n)}{(ab+b^2)/2a}},$$

$$X_a^n = X_b^n + v_0(t_a^n - t_b^n) - \frac{b}{2}(t_a^n - t_b^n)^2 \quad \text{and} \quad v_a^n = v_0 - b(t_a^n - t_b^n) \tag{4.43}$$

4.4.5.3 Free-flow position and time

If any following vehicle $n\,(2 \le n \le N)$ has been identified as having delayed motion, it will have its own free-flow position and time. Under this condition, if $X_b^n \ge \bar{X}_b$, we can use known acceleration variables of t_a^n, X_a^n and v_a^n to find the free-flow position X_v^n and time t_v^n, but if the braking position X_b^n is upstream of $\bar{X}_b$ (out of boundary case), these values can be calculated from the detector position corresponding to the possible departure time $\bar{t}_v^n$ at position $\bar{X}_v$ (see Figure 4.5b).

1) If braking position $X_b^n \ge \bar{X}_b$

$$t_v^n = t_a^n + \frac{v_0 - v_a^n}{a} \quad \text{and} \quad X_v^n = X_a^n + v_a^n(t_v^n - t_a^n) + \frac{a}{2}(t_v^n - t_a^n)^2 \qquad (2 \le n \le N) \tag{4.44}$$

2) If braking position $X_b^n < \bar{X}_b$

In such cases, addition information of estimated speed v_d^n at time t_d^n is required; this is the estimated speed of the vehicle corresponding to the known information of the departure time $\bar{t}_v^n$ when the vehicle crossed the detector. The free-flow time t_v^n can be calculated subject to the following three equations:

$$v_0 = v_d^n + a(t_v^n - t_d^n),$$

$$X_v^n = X_d + v_d^n(t_v^n - t_d^n) + \frac{a}{2}(t_v^n - t_d^n)^2 \quad \text{and}$$

$$\bar{X}_v = X_v^n + v_0(\bar{t}_v^n - t_v^n) \qquad (2 \le n \le N)$$

then, the free-flow time t_v^n and position X_v^n are calculated as follows:

$$t_v^n = t_d^n + \sqrt{\frac{2\left[(X_d - \bar{X}_v) + v_0(\bar{t}_v^n - t_d^n)\right]}{a}} \qquad (2 \le n \le N) \qquad (4.45)$$

and

$$X_v^n = X_d + v_d^n(t_v^n - t_d^n) + \frac{a}{2}(t_v^n - t_d^n)^2 \qquad (2 \le n \le N) \qquad (4.46)$$

where $v_d^n = v_0 - a(t_v^n - t_d^n)$.

4.4.5.4 Crossing time to the stop-line

This is the final information we require for the following vehicle n $(2 \le n \le N)$ trajectory. At the position of the stop-line X_s, the arrival time t_s^n and its speed v_s^n are calculated by comparing the fixed position X_s and varying position X_v^n. If the position X_v^n at which free-flow speed is regained is upstream of the stop-line X_s, then the vehicle will pass X_s with free-flow speed v_0; otherwise, it will pass X_s whilst accelerating. Consequently, the crossing time t_s^n and their speed v_s^n can be calculated by comparing two position variables X_v^n and X_s:

1) If braking position $X_b^n \ge \bar{X}_b$

If $X_v^n \le X_s$, then

$$t_s^n = t_v^n - \frac{X_v^n}{v_0} \quad \text{and} \quad v_s^n = v_0 \qquad (2 \le n \le N) \qquad (4.47)$$

If $X_v^n > X_s$, then

$$t_s^n = t_a^n + \frac{v_s^n - v_a^n}{a} \qquad (2 \le n \le N) \qquad (4.48)$$

where $v_s^n = \sqrt{(v_a^n)^2 - 2aX_a^n}$.

2) If braking position $X_b^n < \overline{X}_b$

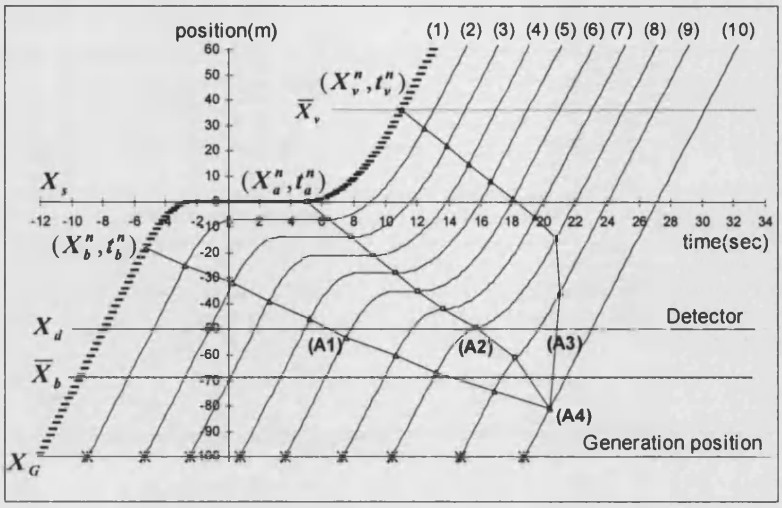
If $X_v^n \le X_s$, then t_s^n is calculated by using (4.47).

If $X_v^n > X_s$, then

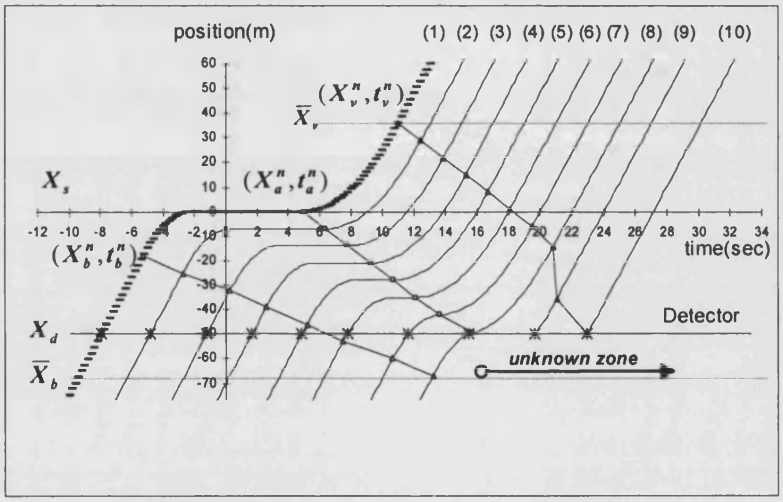
$$t_s^n = t_v^n - \frac{(v_0 - v_s^n)}{a} \qquad (2 \le n \le N) \qquad (4.49)$$

where $v_s^n = \sqrt{(v_d^n)^2 - 2aX_d}$.

Finally, we have a complete set of numerical equations that can estimate arrival times and crossing times (departure times) of any detected vehicles from the detector to the stop-line in respect of the signal displays.



(a) Trajectory from the generation position



(b) Trajectory from the detector position

Figure 4.5 Trajectory of following vehicles

4.4.6 Discussion

As can be seen in Figure 4.5 (a), three different sets of motions have been identified at the position of the detector during one green period. Time and position variables are expanding backwards from the first vehicle's trajectory through the upstream of the detector. In respect of the start of green , any detected vehicles until the time (A1) have crossed the detector with free-flow speed, the times between (A1) and (A2) vehicles have crossed it while braking, the times between (A2) and (A3) vehicles have crossed it while accelerating, and any following vehicles after the time (A4) when

$t_b^n = t_a^n = t_v^n$ will cross the detector with free-flow speed. Presumably, the time (A4) is the queue dissipation time of this stage.

In Figure 4.5 (a), we can see vehicles are generated from the position that is further upstream than the detector, so that it is possible to see all trajectories from X_G to X_s. However, in practice, control decisions are made according to the data obtained from the detector alone. Therefore, we need to define some portion of space zone from the detector that cannot be handled. Figure 4.5 (b) shows that the trajectories of 10-vehicle between X_d and X_s. It is possible to describe the complete motion for vehicles (1)-(8), but afterwards arriving vehicles motion can only be described from the detector.

4.5 SENSITIVITY OF DELAY AT DIFFERENT SIGNAL TIMING PLANS

In this section, the delay difference between the vertical queueing model and the KCS traffic model is discussed. Also, the sensitivity of delay with respect to variations in the start of green time is investigated. For both models, the delay of leading vehicle is determined by the current signal display, and then afterwards delay of following vehicles is dominated by the motion of the leading vehicle. In all cases, the delay is defined as the time difference between the possible departure time and the free-flow travel time through the junction; the vertical delay is calculated at position X_s of the stop-line, but the KCS delay is calculated at position $\bar{X}_v$ where the vehicle regains free-flow speed.

4.5.1 Delay calculation

In the following section 4.5.1.1 and 4.5.1.2, the delay difference between the vertical queueing model and the KCS traffic model is presented in depth.

4.5.1.1 Vertical queueing model delay (vertical delay)

The vertical queueing model in traffic signal control represents all vehicles as having the same travel time before joining a queue. Thus, a queue forms vertically at the stop-line without occupying any space on the link. The delay is identified as the time difference between the queue departure time

and its free-flow travel time to the stop-line. In this model, the departure time of the leading vehicle is assumed to coincide with the *start of the effective green time* (Clayton, 1940; Webster, 1958; Allsop, 1970). Then, the delay for following vehicles are calculated on the basis of calculated headways after the leading vehicle.

1) Delay for the leading vehicle

In order to calculate the delay for the leading vehicle, the first order of task is finding the start of the effective green time $t_g + l_s$, it can be obtained by adding the start lag l_s to the start of green time t_g, where $l_s = v_0 / a - \bar{X}_v / v_0$. As seen in Figure 4.6, there are two cases of delay to be considered according to the relation between the time $t_g + l_s$ and the time $\bar{t}_{d,s}^1$, where $\bar{t}_{d,s}^1$ is the free-flow travel time of the leading vehicle n=1 from the detector to the stop-line ($\bar{t}_{d,s}^1 = t_d^1 + (X_s - X_d)/v_0$). The delay for the vehicle n=1 with respect to the time t_g is expressed as $D_V^1(t_g)$, which are calculated as follows:

- If $(t_g + l_s \le \bar{t}_{d,s}^1)$, then

$$D_V^1(t_g) = 0 \tag{4.50}$$

- If $(t_g + l_s > \bar{t}_{d,s}^1)$, then

$$D_V^1(t_g) = (t_g + l_s) - \bar{t}_{d,s}^1 \tag{4.51}$$

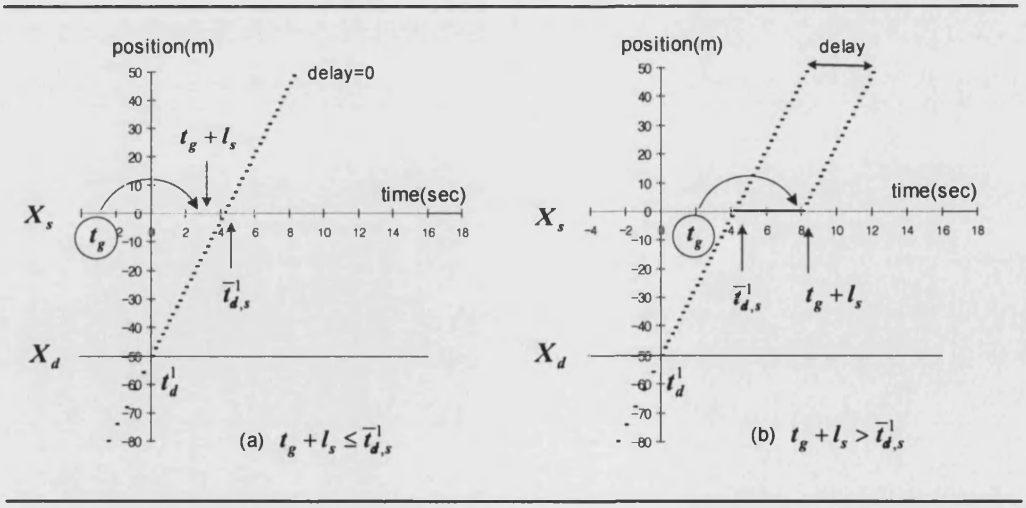


Figure 4.6 Delay for a leading vehicle (Vertical queueing model)

2) Delay for the following vehicles

In order to calculate the delay for the following vehicles $n(2 \le n \le N)$, the first task is finding the possible departure time of them $\bar{t}_s^n = \text{Max}[\hat{t}_s^n, \bar{t}_{d,s}^n]$ at the position X_s, where $\hat{t}_s^n = \bar{t}_s^{n-1} + 1/s$ (s is the saturation departure rate). The vertical delay for the following vehicle n can be calculated as follows:

$$D_V^n = \text{Max}[0, (\bar{t}_s^n - \bar{t}_{d,s}^n)] \qquad (2 \le n \le N) \tag{4.52}$$

where $\bar{t}_{d,s}^n = t_d^n + (X_s - X_d)/v_0$ is the free-flow travel time of the following vehicle n from the detector to the stop-line.

From equations (4.50 - 4.52), the total vertical delay is calculated as follows:

$$D_V = D_V^1(t_g) + \sum_{n=2}^{N} \text{Max}[0, (\bar{t}_s^n - \bar{t}_{d,s}^n)] \tag{4.53}$$

4.5.1.2 KCS traffic model delay (KCS delay)

The delay for the KCS traffic model is calculated on the basis of the start of the green time t_g in accordance with vehicular characteristic variables, such as, acceleration rate, braking rate, free-flow speed and physical queue length.

1) Delay for the leading vehicle

As can see in Figure 4.7, there are three cases of delay we need to consider according to the relation of the time t_g with t_b and t_q. The delay for the leading vehicle n=1 with respect to the time t_g is expressed as $D_K^1(t_g)$, which are calculated as follows:

- If $t_g \le t_b^1$, then

$$D_K^1(t_g) = 0 \tag{4.54}$$

- If $t_b^1 < t_g \le t_q^1$, then

$$D_K^1(t_g) = t_v^1 - t_{d,v}^1 \tag{4.55}$$

where $t_{d,v}^1$ is the free-flow travel time of the leading vehicle n =1 from X_d to X_v^1 : $t_{d,v}^1 = t_d^1 + (X_v^1 - X_d)/v_0$. Using (4.21) and (4.23), we can rewrite equation (4.55) as a function of t_g:

$$D_K^1(t_g) = (t_b^1 - t_d^1) - \frac{(X_b^1 - X_d)}{v_0} + \frac{(ab+b^2)}{2av_0}(t_g - t_b^1)^2 \tag{4.56}$$

- If $t_g > t_q^1$, then

$$D_K^1(t_g) = t_v^1 - t_{d,v}^1 \tag{4.57}$$

using (4.24) and (4.25), we can rewrite equation (4.57) as a function of t_g:

$$D_K^1(t_g) = (t_g - t_d^1) + \frac{v_0}{2a} + \frac{X_d}{v_0} \tag{4.58}$$

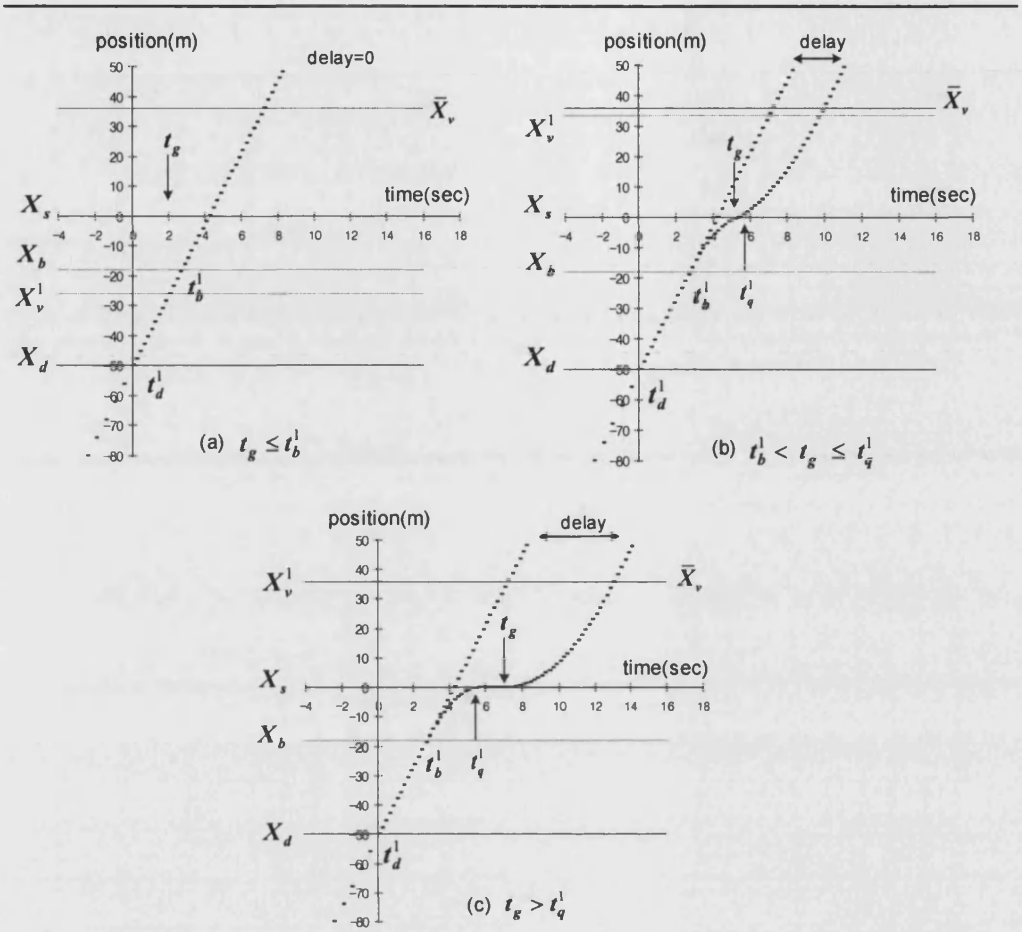


Figure 4.7 Delay for a leading vehicle (KCS traffic model)

2) Delay for the following vehicles

In order to calculate the delay for the following vehicles, the first order of task is finding the possible departure time $\overline{t}_v^n$ for the following vehicles n ($2 \le n \le N$). The time $\overline{t}_v^n$ at the position $\overline{X}_v$ is obtained from the equation (4.33). The KCS delay for the following vehicle n can be calculated as follows:

$$D_K^n = \text{Max}[0, (\overline{t}_v^n - \overline{t}_{d,v}^n)] \qquad (2 \le n \le N) \tag{4.59}$$

where $\overline{t}_{d,v}^n$ is the free-flow travel time of following vehicle n from the X_d to the $\overline{X}_v$ ($\overline{t}_{d,v}^n = t_d^n + (\overline{X}_v - X_d)/v_0$).

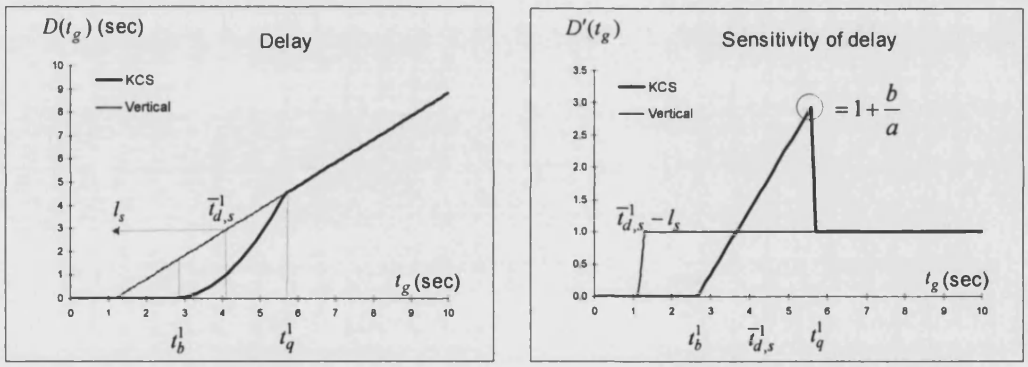
From equations (4.54, 4.55 and 4.57), the total KCS traffic model delay is calculated as follows:

$$D_K = D_K^1(t_g) + \sum_{n=2}^{N} \text{Max}[0, (\bar{t}_v^{\,n} - \bar{t}_{d,v}^{\,n})] \tag{4.60}$$

4.5.2 Sensitivity analysis

4.5.2.1 Sensitivity analysis for the leading vehicle

In this section, the sensitivity of delay to variations in the start of green t_g is discussed. For the leading vehicle, the sensitivity of delay is calculated by differentiating the delay equations (4.50, 4.51, 4.54, 4.56 and 4.58). However, for the following vehicles, the sensitivity of delay to variations in the start of green t_g is calculated by simulation, on the basis of generated vehicles.



(a) Delay (b) Sensitivity of delay

Figure 4.8 Sensitivity of delay for the leading vehicle in varying time t_g

1) Sensitivity analysis for the vertical queueing model

This model gives a simple linear formula for the delay sensitivity. As seen in Figure 4.8 (a), if the green t_g starts before the free-flow travel time minus the start lag l_s, so that $t_g \le \bar{t}_{d,s}^1 - l_s$, no delay is incurred, and the resulting sensitivity is 0. However, from that time if t_g is increased by an amount ε, delay is increased by an identical amount ε, so that the resulting sensitivity is 1. By

differentiating the delay equations (4.50 and 4.51) with respect to the start of green t_g, we can get the sensitivity of delay $D_V^{1\prime}(t_g)$ as follows:

$$D_V^{1\prime}(t_g)=\begin{cases}0 & t_g \le \overline{t}_{d,s}^1 - l_s \\ 1 & t_g > \overline{t}_{d,s}^1 - l_s\end{cases} \tag{4.61}$$

2) Sensitivity analysis for the KCS traffic model

This model gives three different cases of delay sensitivity. If the green t_g starts before the vehicle starts to brake t_b^1, no delay is incurred, and the resulting sensitivity is 0. If the green t_g starts between t_b^1 and t_q^1, the delay increases quadratically, and the resulting sensitivity increases linearly. Finally, from t_q^1 if t_g is increased by an amount ε, delay is increased by an identical amount ε, so that the resulting sensitivity is 1. By differentiating the delay equation (4.54, 4.56 and 4.58) with respect to the varying time t_g, we can get the sensitivity of delay $D_K^{1\prime}(t_g)$ as follows:

- If $t_g \le t_b^1$, then

$$D_K^{1\prime}(t_g)=0 \tag{4.62}$$

- If $t_b^1 < t_g \le t_q^1$, then

$$D_K^{1\prime}(t_g)=\frac{(ab+b^2)}{av_0}(t_g - t_b^1) \tag{4.63}$$

- If $t_g > t_q^1$, then

$$D_K^{1\prime}(t_g)=1 \tag{4.64}$$

3) Maximum sensitivity of delay

The maximum sensitivity of delay for the leading vehicle in the vertical queueing model is always $D_V^{1'}(t_g)=1$ (4.61). In contrast, the maximum sensitivity of delay for the leading vehicle in the KCS traffic model is obtained when $t_g = t_q^1$. By differentiating the equation (4.21), we get

$$\text{Max } D_K^{1'}(t_g) = 1+\frac{b}{a} \tag{4.65}$$

As seen in Figure 4.8, delay $D(t_g)$ and sensitivity $D'(t_g)$ for the leading vehicle n=1 have been evaluated with variations of t_g ranging from 0 to 10 sec. In this example, we suppose that the leading vehicle is detected at time t_d=0, and with given condition parameters of X_d = -50 m, X_s = 0, acceleration rate a = 2 m/sec^2, braking rate b = 4 m/sec^2 and free-flow speed v_0 = 12 m/s. Other calculated position variables are: the braking time t_b^1 = 2.67 sec (4.9), the stopping time t_q^1 = 5.67 sec (4.13), the free-flow travel time $\overline{t}_{d,s}^1$ = 4.17 sec from the detector to the stop-line, and the start lag, which is about 3 sec. For this example the maximum sensitivity of delay is $D_K^{1'}(t_g)$=3, and it incurs just before the time t_g = 5.67 sec.

4) Summary

Under the condition of the braking rate is greater than the acceleration rate ($b \geq a$), the sensitivity of delay between the vertical queueing model and the KCS traffic model with respect to t_g is summarised as follows:

- The vertical delay starts at time which $t_g = \overline{t}_{d,s}^1 - l_s$ from that time the delay increases linearly.
- In contrast, the KCS traffic model delay starts at time which $t_g = t_b^1$ from that time until $t_g = t_q^1$ the delay increases quadratically and thereafter it increases linearly.
- The delay in KCS traffic model starts about time of $t_b^1 - (\overline{t}_{d,s}^1 - l_s)$ later
- The delay of two control models become equal at time which $t_g = t_q^1$.
- The delays differ only if t_g lies between $\overline{t}_{d,s}^1 - l_s$ and t_q^1.

- If t_g lies between $\overline{t}_{d,s}^1 - l_s$ and $\overline{t}_{d,s}^1$, the vertical queueing gives higher sensitivity than the KCS traffic model.
- If t_g lies between $\overline{t}_{d,s}^1$ and t_q^1, the KCS traffic model shows the highest sensitivity of delay which is equal to $D_K^{1\prime}(t_g = t_q^1) = 1 + b/a$, but the vertical queueing is $D_K^{1\prime}(t_g = t_q^1) = 1$.

4.5.2.2 Sensitivity analysis for many vehicles

In this section, the following results are based on the simulation. Two different cases of traffic are considered: a low density case and a high density case. In the present examples, the maximum number of vehicles that can be held in the queue with this given condition is eight vehicles. The detector is located X_d = -50m from the stop-line, and the minimum spacing of each following vehicle is L = 7m, saturation departure time is 1.8 sec/vehicle (2,000 vehicles/hour) and all other conditions are same as the single vehicle case. In the following examples, two cases of vehicles are generated based on *shifted exponential distribution of headways H*, which is given by

$$H = h_0 - \frac{1}{\alpha}\ln(u) \tag{4.66}$$

where

u is the random value generation, in which variables are generated with equal probability between [0, 1],

h_0 is the minimum gap of following ($h_0 \geq \tau + L/v_0$),

α is the density parameter (the bigger α generates greater flow),

τ is the reaction time.

Using equation (4.66), we can generate the arrival time of the detector without causing any headway violation. The time t_d^n can be given by

$$t_d^n = t_d^{n-1} + H \qquad (n \geq 2) \tag{4.67}$$

Starting from the first vehicle's detection time t_d^1=0 sec, the following seven vehicles are generated by using (4.67). The generated detector times for the low density case and the high density case are shown in Table 4.1.

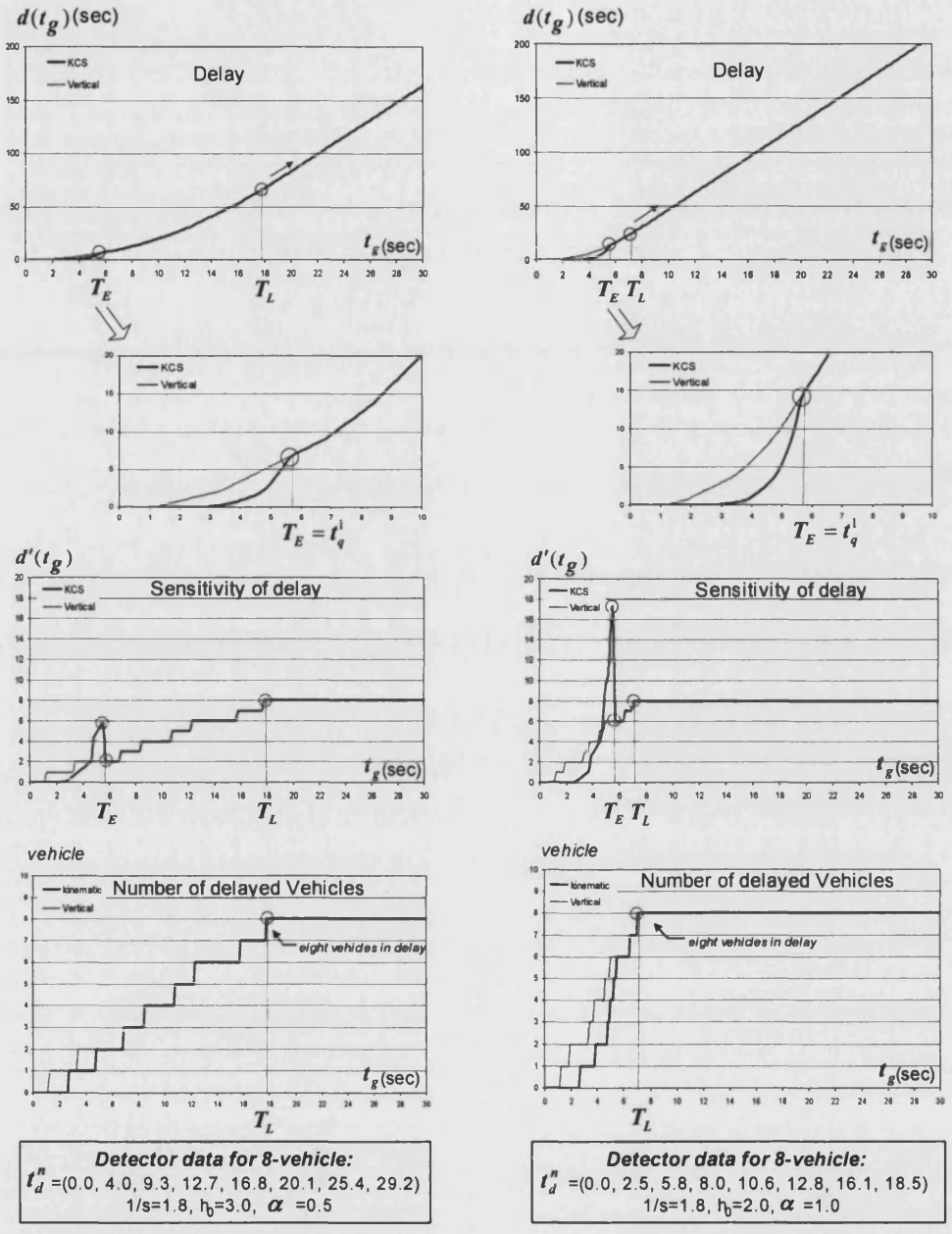
Table 4.1 Detector time generation: using shifted exponential distribution of headways

		Vehicle n=							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Low density α=0.5 h_0=3.0	u	-	0.613	0.321	0.824	0.569	0.851	0.312	0.67
	h_0	-	3.97	5.27	3.38	4.12	3.32	5.33	3.79
	t_d^n	0.0	4.0	9.3	12.7	16.8	20.1	25.4	29.2
High density α=0.9 h_0=2.0	u	-	0.613	0.321	0.824	0.569	0.851	0.311	0.671
	h_0	-	2.54	3.26	2.21	2.62	2.18	3.29	2.44
	t_d^n	0.0	2.5	5.8	8.0	10.6	12.8	16.1	18.5

Based on Table 4.1 data, the sensitivity of delay for eight vehicles are tested with respect to the variations in the start of green time t_g: in the range of t_g = 0 ~ 30 sec and the t_g is incremented by 0.1 sec. The resulting delay, sensitivity and number of delayed vehicles are shown in Figure 4.9.

As noted in Section 4.5.1.1, the vertical delay starts at time which $t_g = \bar{t}_{d,s}^1 - l_s$ and the KCS delay starts at time which $t_g = t_b^1$. Finally, they become equal at time which $t_g = t_q^1$. As can be seen in Figure 4.9, let T_E be the time at which the total delay of eight vehicles for the KCS traffic model and the vertical queueing model become same, namely $T_E = t_q^1$. In the same sense T_L is the time at which total delay becomes increasing linearly. If t_g starts before T_L, which means that all eight vehicles have not delayed yet; in this range, we can suppose that some vehicles may go through the stop-line without experiencing any delay, thus the sensitivity fluctuates. However, if t_g starts after T_L, which means that all eight vehicles have been delayed, the sensitivity of delay in this range with respect to t_g is equal to the total number of delayed vehicles.

The sensitivity of delay for the vertical queueing model is equal to the number of delayed vehicles in the time range. But it differs for the KCS traffic model if the t_g starts before T_E, in some time range it shows a sensitivity higher than the total number of delayed vehicles. The simulation results for the low density case and the high density case are discussed below:



(a) Low density case

(b) High density case

Figure 4.9 Sensitivity of delay for the following vehicles in varying time t_g

1) For the low density case simulation (see Figure 4.9 (a))

- The vertical delay is greater than or equal to the KCS delay.
- T_E = 5.7 sec and T_L = 17.8 sec.
- At time t_g = 5.5 sec, the KCS traffic model shows the sensitivity of delay $D_K^{'}$ = 5.76, and the vertical queueing shows $D_V^{'}$ = 2, which is equivalent to numbers of delayed vehicles.
- At time t_g = 17.8 sec, the sensitivity of delay for both models is equal to 8. From that time, we can see that all vehicles will experience some delay.

2) For the high density case simulation (see Figure 4.9 (b))

- The vertical delay is greater than or equal to the KCS delay.
- T_E =5.7 sec and T_L =7.1 sec.
- At time t_g = 5.5 sec, the KCS traffic model shows a sensitivity of delay $D_K^{'}$ = 17.28, and the vertical queueing model shows $D_V^{'}$ = 6, which is equivalent to numbers of delayed vehicles. At this time, the sensitivity of delay for the KCS traffic model is about three times greater than for the vertical queueing model.
- At time t_g = 7.1 sec, the sensitivity of delay for both models is equal to 8. From that time, we can see that all vehicles will experience some delay.

4.6 DISCUSSION

In this chapter, a Kinematic Car-following model at Signalised junctions (the KCS traffic model) was proposed. As explored in this chapter, the delay estimated in the vertical queueing model and the KCS traffic model differs around the time of the green start, but so long as there is a queue the

delay estimates from these models are identical. With respect to variations in the start of green time t_g, and provided that the braking rate b exceeds the acceleration rate a, so that $b \geq a$, the delay calculated in the vertical queueing model is greater than or equal to the KCS traffic model; however, the maximum sensitivity of delay in the KCS traffic model is greater than that in the vertical queueing model.

In the vertical queueing model, the delay is calculated according to the start of effective green time, so that when the effective green time starts before the free-flow travel time, no delay is caused. In contrast, in the KCS traffic model, the delay is calculated according to relations between the trajectory of the vehicle and the start of green time, so that when the t_g occurs before the vehicle starts to brake, no delay is caused. Thus, with respect to the time of detection of the leading vehicle, the delay difference between these two models only occurs when the t_g falls within the time rage of the free-flow travel time minus the start lag and the time at which the vehicle stops completely after full braking. In this time range, the vertical queue delay is greater than that the KCS one, but if t_g falls before the free-flow travel time, the vertical queueing model shows greater sensitivity than the KCS traffic model in relation to variation in t_g. From these results, it is clear that the leading vehicle trajectory is an important determinant for calculating the delay at signalised junctions.

The vertical queuing model does not consider any braking motion other than an abrupt halt at the stop-line, so that it assumes all vehicles have the same travel time before joining the queue at the stop-line. Using the start lag adjustment, the departure time of the leading vehicle at the stop-line is assumed to coincide with the start of the effective green time, which is the reason why the delay starts before that in the KCS traffic model, and then the following vehicles depart the stop-line with the same headways. In this case, a queue delay can be calculated at the position of the stop-line.

By contrast, the KCS traffic model calculates delay on the basis of each individual vehicle's trajectory with respect to the start of green time as well as vehicular characteristics, such as acceleration rate, braking rate and free-flow speed. Thus, the departure time of vehicles from the stop-line can not be assumed beforehand, but rather will depend on how much braking the vehicle undergoes. In this case, an approach delay incurred from the upstream braking position to the downstream free-flow position through the stop-line can be calculated in detail. Under these different model characteristics, the departure headways (or the times between crossing the stop-line) of successive vehicles are not the same from the beginning, but converge gradually to the saturation departure time if the green period is long enough and there are sufficient numbers of vehicles.

As noted, the motion of a leading vehicle at the signalised junction is affected by the current signal display, and the motion of following vehicles is dominated by that of the vehicle in front (the previous vehicle). Once one vehicle has passed through the junction without delay, then we can suppose that any further vehicles either in the same stage or in the same green period can also go through without incurring delay. However, if the leading vehicle has passed the junction with some delay, the following vehicles may or may not experience delay, depending on their arrival pattern. The formulae established in this chapter will be used as part of the traffic model in Chapter 5.

5. A DYNAMIC OPTIMISATION METHOD

5.1 INTRODUCTION

This chapter developed a systematic optimising method for dynamic signal control in stage-based that requires a real-rime detector data and a set of stage sequences. This optimising method consists of two parts: a *dynamic optimiser* and a *traffic model*. They are processed in alternating order, in which the dynamic optimiser provides future signal timings to the traffic model, and the traffic model subsequently provides estimates of operational performance back to the dynamic optimiser. In this way, the dynamic optimiser can make the control decisions, whether 'roll-forward' (i.e. extension of the current stage is beneficial) or 'stage change' (i.e. change to the next stage is beneficial). The whole optimisation process is established to conform Gartner's (1983) rolling-horizon concept. In cases where the roll-forward period is shorter than the detection period, the detector data will be used several times over.

In the optimising process, a plan is developed and evaluated for a certain lookahead period. The duration that is adopted initially corresponds to one complete cycle and is the same for all streams at the junction. Starting from the end of each stage minimum green, the dynamic optimiser explores all possible combinations of the future signal timings according to an *exhaustive search method.* The search range for each stage is defined between the minimum and the maximum permissible green durations. During this exploration, an objective function that is used to estimate operational performance based on the current state of traffic and the signal timings identified by the dynamic optimiser. This objective function is based on the KCS traffic model that was introduced in Chapter 4.

The control decision to extend the current stage is considered at regular intervals based on the criterion of minimising the total rate of delay in all streams at the junction. At the time of optimisation, two delays are compared, corresponding to the minimum achievable if the current stage ends immediately and the minimum achievable if the current stage is extended by at least one time-step. By fixing the current stage, process the extension tests for the following stages until the maximum green is reached, and the lowest delay which is denoted as D^* is stored by the dynamic optimiser. If any extension to the current stage gives less delay than D^*, the exhaustive search is terminated and the dynamic optimiser extends the current stage, rolls forward by one time-step and repeats the optimisation (see Figure 5.1 and 5.3).

The objective function used for performance evaluation consists of three distinct components of delay: detection period delay, prediction period delay and terminal cost. The first of these is the delay incurred by the vehicles that have been detected, which is calculated by using the chosen traffic model. The second is an estimate of delay that will be incurred by vehicles that have not yet been detected, which is calculated based on the trajectory of the last detected vehicle and future mean arrivals during the lookahead period: the vertical queueing concept is applied to this traffic. Any residual queues at the end of this period are represented through the terminal cost function that was suggested by Robertson and Bretherton (1974).

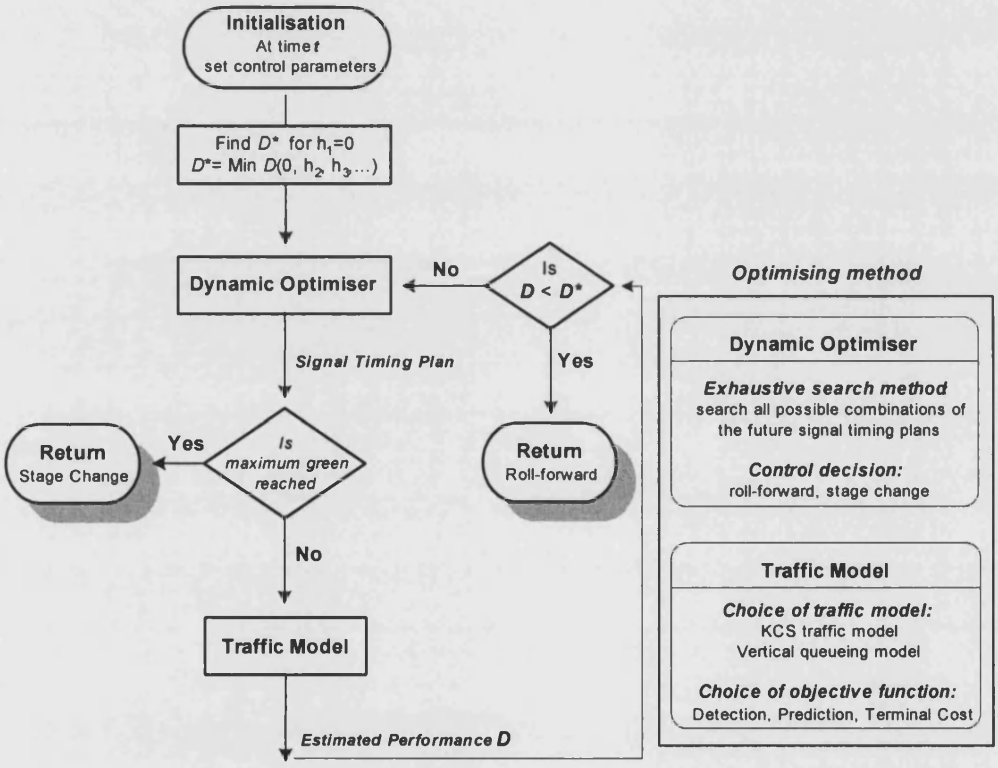


Figure 5.1 Flowchart of the decision making process

5.2 NOTATION

The following notation will be used throughout this chapter, and additional notation is introduced in the following section.

Let

t be the time at which the optimisation takes place; this is started initially at the end of the first stage minimum green,

Δh be the roll-forward period,

m be the number of stages in the cycle,

k be the number of streams at the junction ,

t_j^* be the time at which the first uniform arrival reaches the stop-line that is the earliest undelayed travel time between the detector and the stop-line at time t. From that time the uniform queue would start to build up, $t_j^* = t - (X_d / v_0)$ $(1 \le j \le k)$,

$\hat{C}$ be the end of the lookahead period (planning period) which is defined as the time up to the end of the next occurrence of the stage that is currently running,

$\phi\hat{C}$ be the duration of the lookahead period of about one-cycle that is defined as the time difference between the beginning of the current stage and $\hat{C}$,

q_j be the mean arrival rate for stream j at $X_d (1 \le j \le k)$,

s_j be the saturation departure rate for stream j at $\bar{X}_v (1 \le j \le k)$,

5.3 OPTIMISING PROCESS

At the time of each optimisation, the end of the lookahead period $\hat{C}$ is defined as the time up to the end of the next occurrence of the stage that is currently running. In the meanwhile, the duration of the lookahead period of about one-cycle $\phi\hat{C}$ is defined from the beginning of the current stage and up to the end of $\hat{C}$. In the decision making process, the objective function calculates the mean rate of delay incurred during the $\phi\hat{C}$ plus the terminal cost based on the exhaustive search method. During the extension tests, two achievable minimum delays are calculated and compared in each optimising process: the minimum delay obtained by fixing the current stage and the minimum delay obtained by extending the current stage. If there is any possibility of obtaining the lower delay by

extending the current stage, then the current stage is extended by a roll-forward period by Δh seconds; otherwise a change of stage is initiated.

5.3.1 Application of exhaustive search method to the optimising process

The lookahead period is adopted initially corresponding to one complete set of the minimum cycle. As seen in Figure 5.2, when the green has started for the stream B, we suppose that it will remain green until the minimum green for the stage has elapsed, that is the shortest period of green time given to any stage in the signalised junction. In cyclic order, the red follows after the inter-stage time, which includes the duration of the amber and the red and amber; sometimes all red period is included (depending on the junction configurations). In the UK, the normal minimum green is 7 seconds, and the inter-stage is 5 seconds of which the amber is 3 seconds and the red and amber is 2 seconds. In stage-based traffic-responsive signal control, sum of the stage minimum green and inter-stage time is referred to as a minimal one-cycle period.

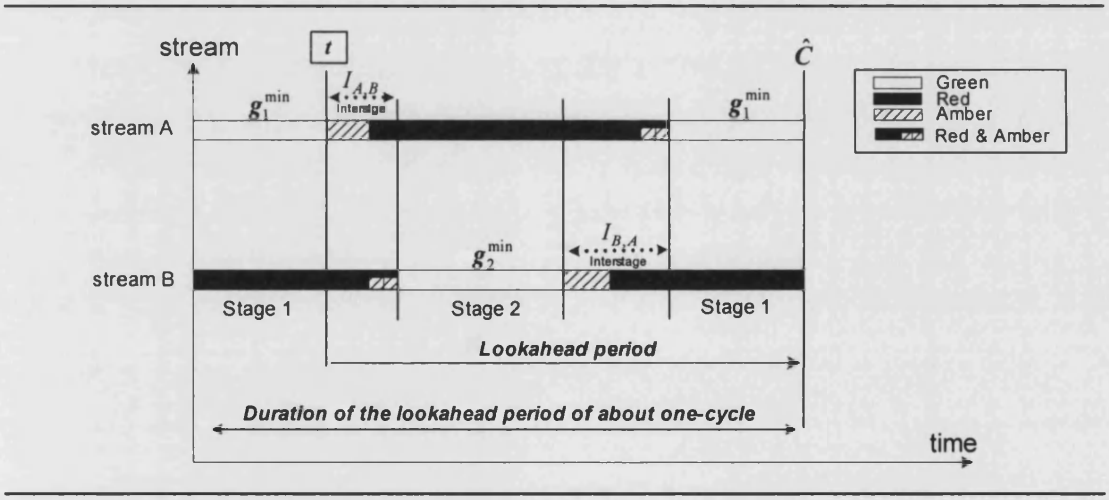


Figure 5.2 Minimum green and red

At time t, the lookahead period varies in relation to the duration of search period, in which the dynamic optimiser tests all possible combinations of the future signal timing plans within a given boundary of the searching period. In this way, the dynamic optimiser tests all possible extensions of the stage sequences, including the stage that is currently running and the stages after following that. This is also referred to as a depth-first search method because all possible extension trees are searched from the planned ending stage to the current stage (backtracking) before another new path

is considered. A separate estimation is made for the delay in each stream of traffic, but control decisions are made based on the summed total rate of delay incurred over all streams at the junction during the lookahead period of about one-cycle.

As shown in Figure 5.3, the extension test starts initially at the end of each following stage minimum. In each iteration, the end of the lookahead period $\hat{C}$ is defined as the time up to the end of the next occurrence of the stage that is currently running. However, the duration of the lookahead period of about one-cycle $\phi\hat{C}$ includes the duration of the current stage and the duration of each of the stages in the whole sequence. At the time of optimisation, two delays are compared, corresponding to the minimum achievable if the current stage ends immediately and the minimum achievable if the current stage is extended by at least one time-step. If there is any possibility of obtaining the lower delay by extending the current stage, then the current stage is extended by one roll-forward period Δh; otherwise, a change of stage is initiated.

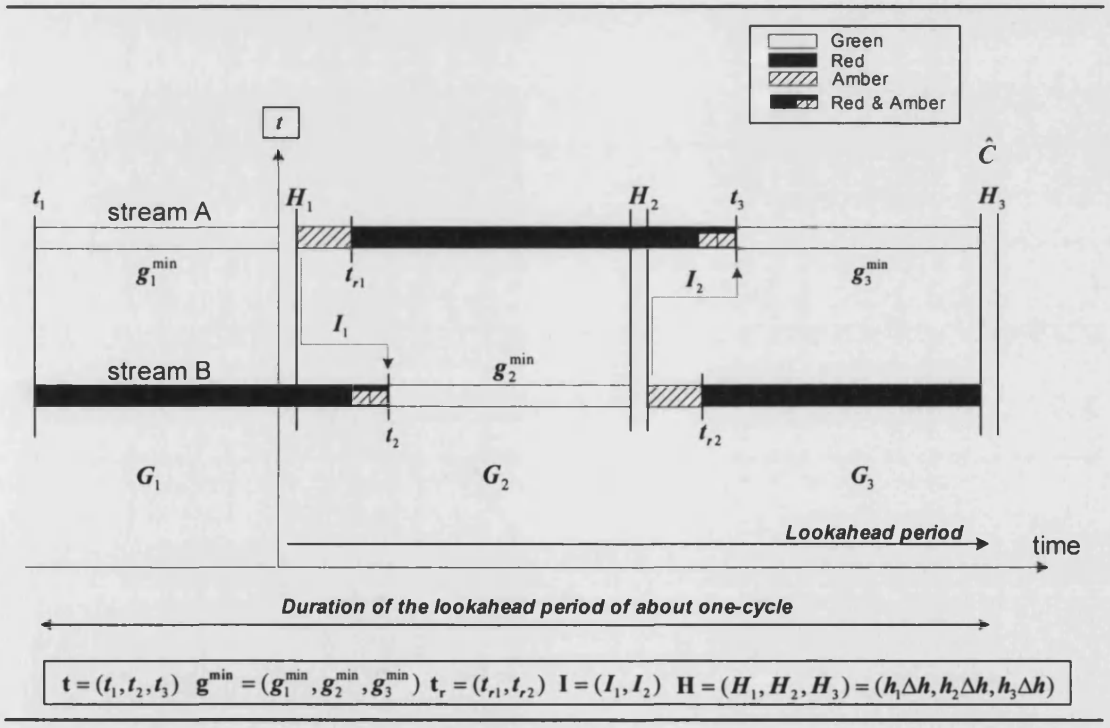


Figure 5.3 Lookahead period calculation at time t (for the junction has two stages)

Let

$\mathbf{G}$ = $(G_1, G_2, .., G_{m+1})$ be a set of stage sequences to be initialised before processing the extension test. It is specified based on the number of stages at the junction,

$\mathbf{h}$ = $(h_1, h_2, .., h_{m+1})$ be a set of the integer-valued control parameters,

$\mathbf{H}$ = $(H_1, H_2, .., H_{m+1}) = (h_1\Delta h, h_2\Delta h, .., h_{m+1}\Delta h)$ be a set of the extension parameters,

$\mathbf{g^{min}}$ = $(g_1^{min}, g_2^{min}, .., g_{m+1}^{min})$ be a set of the minimum green times,

$\mathbf{g^{max}}$ = $(g_1^{max}, g_2^{max}, .., g_{m+1}^{max})$ be a set of the minimum green times,

$\mathbf{I}$ = $(I_1, I_2, .., I_i)$ be a set of the inter-stage times $(i = m)$,

$\mathbf{A}$ = $(A_1, A_2, .., A_i)$ be a set of the amber times $(i = m)$,

$\mathbf{t}$ = $(t_1, t_2, .., t_{m+1})$ be a set of the green indications that can be calculated in relation to the set of $\mathbf{H}$, $\mathbf{g^{min}}$ and $\mathbf{I}$, where t_i is the beginning time of each stage,

$\mathbf{t_r}$ = $(t_{r1}, t_{r2}, .., t_{rm+1})$ be a set of the red indications, where t_{ri} is the beginning time of each red that follows t_i,

D be the total rate of delay over all streams at the junction that is incurred by the set of extension parameters $\mathbf{H}$ and the resulting duration $\phi\hat{C}$,

T_{S1} be the duration of the remaining green for the current stage at time t, which is calculated corresponding to the difference between the current time of optimisation and the time at which maximum green will elapse,

T_{S2} be the duration of the search period for the following stages at time t, which is calculated corresponding to the difference between minimum green and maximum green.

As seen in Figure 5.3, the first task is to specify a set of the stage sequences $\mathbf{G}$, in which stages are numbered in order and they are required to occur sequentially after each following inter-stage time has elapsed. Respectively, a set of beginning time for $\mathbf{G}$ is denoted as $\mathbf{t}$ in which the beginning time of the following stages in the whole sequence is specified. For the two stage case, $\mathbf{G}$ = (G_1, G_2, G_3) and respectively $\mathbf{t} = (t_1, t_2, t_3)$. Hence, at time t, the end of the lookahead period $\hat{C}$ is calculated based on the $\mathbf{t}$ with respect to the extension parameter $\mathbf{H}$, where $\mathbf{H}$ is determined by the set of integer-valued control parameter $\mathbf{h}$.

The optimisation starts initially at time $t = t_1 + g_1^{\min}$. From that time t, a set of the green indications $\mathbf{t}$ and the red indications $\mathbf{t}_r$ with respect to the extension parameter $\mathbf{H}$ can be calculated sequentially as follows:

$$\mathbf{t} = (t_1,\ t_2 = t_1 + g_1^{\min} + H_1 + I_1, \ldots\ldots, t_i = t_{i-1} + g_{i-1}^{\min} + H_{i-1} + I_{i-1}) \quad (i = m+1) \qquad (5.1)$$

$$\mathbf{t}_r = (t_{r1} = t_1 + g_1^{\min} + H_1 + A_1, \ldots\ldots,\ t_{ri} = t_i + g_i^{\min} + H_i + A_i) \quad (i = m) \qquad (5.2)$$

using (5.1), the end of the lookahead period $\hat{C}$ at time t is calculated by

$$\hat{C} = t_i + g_i^{\min} + H_i \qquad (i = m+1) \qquad (5.3)$$

and the duration of the lookahead period of about one-cycle $\phi\hat{C}$ is given by

$$\phi\hat{C} = \hat{C} - t_1 \qquad (5.4)$$

• Exhaustive search algorithm in the dynamic optimiser:

Step 0: (Initialisation)

Set the time sequences $\mathbf{t}$ (5.1) with respect to the number of stages m at the junction.

The initial values of the control parameters are $\mathbf{h}^* = (0, 0, \ldots, h_i = 0)$. Thus, the initial set of the extension parameters become $\mathbf{H}^* = (0, 0, \ldots, H_i = 0)$, where $i = m+1$.

By fixing $t_1 = 0$, the optimisation time will be at time $t = t_1 + g_1^{\min}$.

The initial delay D is set to $D^* = (\mathbf{H}^*, \phi\hat{C}) = \infty$.

Step 2: (Exhaustive search algorithm: stage extension test)

Using the equations (5.1) and (5.4), update $\mathbf{t}$ and $\phi\hat{C}$ with respect to $\mathbf{H}$.

Two minimum achievable delays are compared in this step; if the current stage is fixed, and if the current stage is extended at least by one time-step. By fixing the current stage, process the extension tests for the following stages until the maximum green is reached, and the lowest delay which is denoted as D^*, is stored by the dynamic optimiser. If any extension to

the current stage gives less delay than D^*, the exhaustive search is terminated and the dynamic optimiser extends the current stage by Δh; otherwise, a change of stage is initiated.

Step 2-1) At time t, search all possible extensions until the maximum green is reached. The lowest delay D^* associated with the control parameters $\mathbf{H}^*$ is stored in the dynamic optimiser in this step.

For (h_1=0) // Exhaustive test for $H_1 = h_1 \Delta h$=0

{

For (h_2=0; $h_2 \le T_{S2}$; Δh++)// Exhaustive test for $H_i = h_i \Delta h$ $(2 \le i \le m+1)$

{

For (h_3=0; $h_3 \le T_{S2}$; Δh++)

⋮

For (h_i=0; $h_i \le T_{S2}$; Δh++)

$\mathbf{H} = (0, h_2 \Delta h, .., h_i \Delta h)$, where $i = m+1$

$D = D(\mathbf{H}, \phi \hat{C})$

If ($D < D^*$), then $D^* = D$

} End for the following stages

} End for H_1=0

Step 2-2) At time $t = t + \Delta h$, process the extension tests for the following stages, including the current stage, until the maximum green is reached. During this, if any extension gives less delay than step 2-1, the exhaustive search is terminated and the dynamic optimiser returns 'roll-forward'; otherwise, a change of stage is initiated.

For (h_1=1 $h_1 \le T_{S1}$; Δh++) // Exhaustive test for $H_1 = h_1 \Delta h$

{

For (h_2=0; $h_2 \le T_{S2}$; Δh++) // Exhaustive test for $H_i = h_i \Delta h$ $(2 \le i \le m+1)$

{

For (h_3=0; $h_3 \le T_{S2}$; Δh++)

$\vdots$

For (h_i=0; $h_i \le T_{S2}$; Δh++)

$\mathbf{H} = (h_1\Delta h, h_2\Delta h, .., h_i\Delta h)$, where $i = m+1$

$D = D(\mathbf{H}, \phi\hat{C})$

If ($D < D^*$), then $D^* = D$

Return ('Roll-forward the current stage') $\Rightarrow$ go to Step 3

} End for the following stages

} Ends for H_1 extensions

Return ('Change to the new stage') $\Rightarrow$ go to Step 4

Step 3: (Roll-forward)

If the control decision is 'roll-forward the current stage' which means that it is beneficial to extend the current stage by Δh. The next optimisation time will be $t + \Delta h$.

$\Rightarrow$ go to Step 2 with $t = t + \Delta h$ and $t_1 = t_1$.

Step 4: (Change to the new stage)

If the decision is 'change to the new stage', which means that no benefit is anticipated in extending the current stage. The next optimisation time will begin after the inter-stage time for the current stage plus the minimum green time for the next stage, which will be $t_2 + g_2^{min}$.

$\Rightarrow$ go to Step 2 with $t = t_2 + g_2^{min}$ and $t_1 = t_2$.

5.4 DELAY CALCULATION FOR THE LOOKAHEAD PERIOD OF ABOUT ONE-CYCLE

5.4.1 Detection period delay

This is the delay incurred by vehicles that have been detected but have not yet crossed the stop-line at the time of the optimisation. These are the vehicles currently in the stream between the detector

and the stop-line at time t, which is denoted as $\bar{N}_j$ $(1 \le j \le k)$ and they will be used for the stage extension test. In each iteration, the KCS traffic model estimates the time of each vehicle crossing the stop-line and what delay will be involved in relation to the set of green indications $\mathbf{t}$. Among them, if any vehicles cannot pass the stop-line at the end of the lookahead period $\hat{C}$ will be included in the terminal cost function.

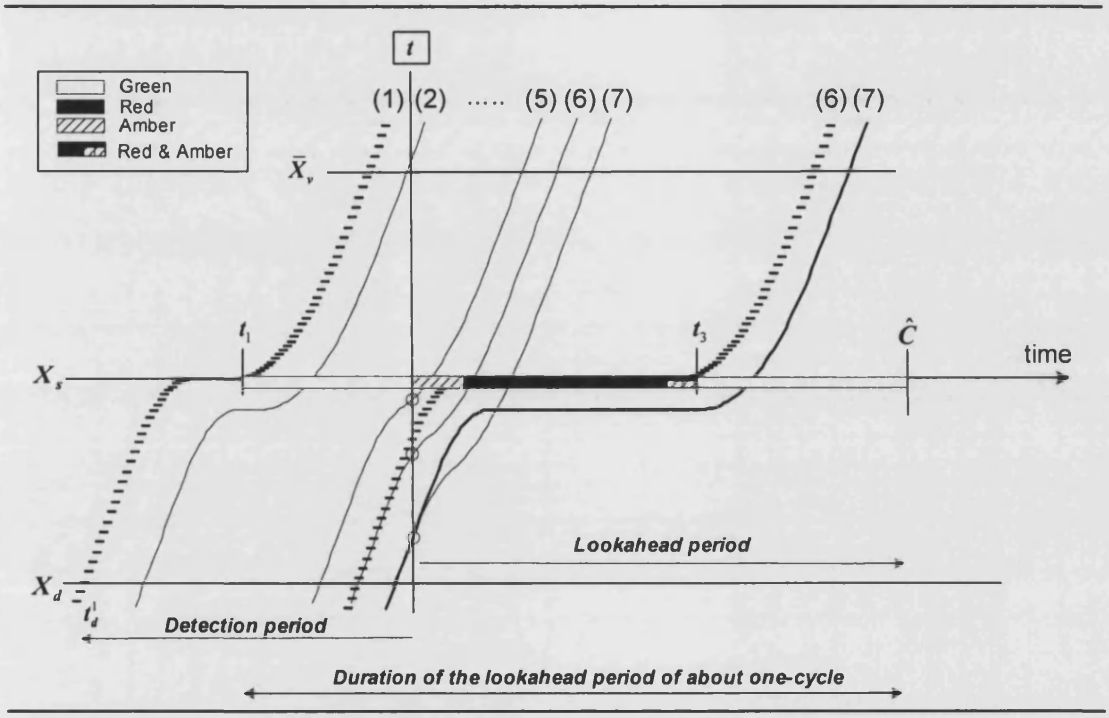


Figure 5.4 Detection period delay calculation in relation to $\mathbf{t}$

All detected vehicles in each stream are numbered in order by giving 1 for the leading vehicle and 2, 3,.., N for the following vehicles, and they are required to depart sequentially according to the green time being considered. At time t, based on the information provided from the detector, the controller identifies the stage leading vehicle in each stream, and also the first vehicle and the last vehicle between the detector and the stop-line. As long as the current stage continues, the stage leading vehicle remains the same, but in successive optimisation processes, the first vehicle and the last vehicle in each stream change according to departures and new arrivals.

In the example shown in Figure 5.4, there are three following vehicles between the detector and the stop-line (the vehicles 5, 6 and 7), so these are the vehicles to be considered in the stage extension

test at time t. Among them, vehicle (5) is identified as first vehicle and vehicle (7) is identified as last vehicle. By taking t_1 as the start of the green time in the KCS traffic model, only vehicle (5) can cross the stop-line during the current green. Hence, we need an additional trajectory estimation for vehicle (6) and vehicle (7) by supposing that they will take t_3 as the start of the green time. Consequently, if the stage changes before vehicles (6) and (7) cross the stop-line, vehicle (6) will be identified as the stage leading vehicle as well as the fist vehicle in the next optimisation process. In the rolling-horizon process, vehicles that have already been considered in the previous optimisation but were unable to pass the stop-line at that time will be considered again in the next optimisation.

In addition, the duration of the detection period for a stream is calculated according to the vehicles that have been detected at time t. This is defined as the maximum interval between the vehicles being detected and crossing the stop-line, in which the detection period starts at time of the stage leading vehicle t_d^1 crossing the detector and ends at time t. In contrast, the duration of the prediction period starts at time t and continues until the end of the lookahead period $\hat{C}$.

At time t, each stream $j\,(1 \le j \le k)$ may have a certain number of vehicles $\bar{N}_j$ between the detector and the stop-line. These are the vehicles to be used for the stage extension test. If $\bar{N}_j$=0 is encountered which means that there are no vehicles between the detector and the stop-line, so that the detection period delay for the stream j is 0. In each optimisation, the delays incurred in all streams are summed together, but for the terminal cost calculation the number of residual vehicles at time $\hat{C}$ has to be summed separately for the streams are currently green or red.

In the optimisation process, the total detection period delay over all streams at the junction with respect to **t** is calculated by

$$D_1 = \sum_{j=1}^{k} \sum_{n=1}^{\bar{N}} d_j^n \tag{5.5}$$

where $d_j^n = (t_{vj}^n - \bar{t}_{vj}^n)$ is the delay incurred by vehicle n in stream $j\,(1 \le j \le k, 1 \le n \le \bar{N})$, in which t_{vj}^n is the arrival time of the vehicle n from X_d to $\bar{X}_v$ and $\bar{t}_{vj}^n$ is the free-flow travel time of the vehicle n from X_d to $\bar{X}_v$.

The total number of residual vehicles at time $\hat{C}$ is counted by comparing the estimated crossing time t_{sj}^{n} and $\hat{C}$ ($1 \le j \le k$, $1 \le n \le \bar{N}$). Two separate counting is required, when the stream is currently on green and the stream is currently on red:

If the streams are currently green and $t_{sj}^{n} > \hat{C}$, then

$$\bar{N}_0^g ++ \tag{5.6}$$

If the streams are currently red and $t_{sj}^{n} > \hat{C}$, then

$$\bar{N}_0^r ++ \tag{5.7}$$

5.4.2 Prediction period delay

This is an estimate of the delay incurred by vehicles that have not yet been detected at the time of optimisation. This delay can be estimated corresponding to the trajectory of the last detected vehicle and the future mean arrivals during the lookahead period $[t,\ t+\hat{C}]$. In this section, we propose a systematic way of estimating the prediction period delay by combining with the KCS traffic model and the vertical queueing concept. The numerical formulae for the prediction period delay are derived separately for the streams that are currently green and for those are currently red.

Let

N_j be the last serial number of the detected vehicle for stream j at time t, ($1 \le j \le k$),

t_{dj}^{N} be the time of the last detected vehicle N_j crossing the detector X_d ($1 \le j \le k$),

t_{sj}^{N} be the time of the last detected vehicle N_j crossing the stop-line X_s, that is equal to the end time of the detection period for stream j ($1 \le j \le k$),

t_{vj}^{N} be the time of the last detected vehicle N_j crossing the upstream free-flow position $\bar{X}_v$ ($1 \le j \le k$),

t_{0j} be the earliest time of the uniform queue dissipation begin in stream j ($1 \le j \le k$), which is calculated from the trajectory of the last detected vehicle: $t_{0j} = t_{vj}^{N} + 1/s - (\bar{X}_v - X_s)/v_0$,

$Q_j(t_{0j})$ be the number of vehicles in the queue at time t_{0j} in stream j ($1 \le j \le k$),

$\theta_j t_{0j}$ be the time when the cumulative queue $Q_j(t_{0j})$ terminates ($1 \le j \le k$),

t_i be the beginning time of the green indications in the set of $\mathbf{t}$ ($1 \le i \le m+1$),

$Q_j(t_i)$ be the number of vehicles in the queue at time t_i in stream j ($1 \le i \le m+1, 1 \le j \le k$),

$\theta_j t_i$ be the time when the cumulative queue $Q_j(t_i)$ terminates ($1 \le i \le m+1, 1 \le j \le k$),

t_{ri} be the beginning time of the red that follows t_i ($1 \le i \le m$),

$Q_j(t_{ri})$ be the number of vehicles in the queue at time t_{ri} in stream j ($1 \le i \le m, 1 \le j \le k$),

$\theta_j t_{ri}$ be the time when the cumulative queue $Q_j(t_{ri})$ terminates ($1 \le i \le m, 1 \le i \le m$),

$Q_j(\hat{C})$ be the number of vehicles in the queue at the end of the lookahead period in stream j ($1 \le j \le k$).

5.4.2.1 Queue clearance time calculation

In practice, the green duration is supposed to be long enough to serve all vehicles that have already been detected, which is called a queue clearance time for the detected vehicles. In contrast, the queue clearance time in the prediction period is estimated in accordance with the arrival time of the last detected vehicle at the stop-line plus the expected future arrival of vehicles that have not yet reached the detector: this is the planned duration of green needed to clear all possible queues built up between the detector and the stop-line. In general, queueing starts at the beginning of red and departures start at the beginning of green, so that the queue grows during the red and dissipates during the green. By integrating the KCS traffic model with prediction arrival modules, the queue formation and dissipation during the prediction period differs from general queueing concepts; even during the green period, the queue may grow instead of dissipating.

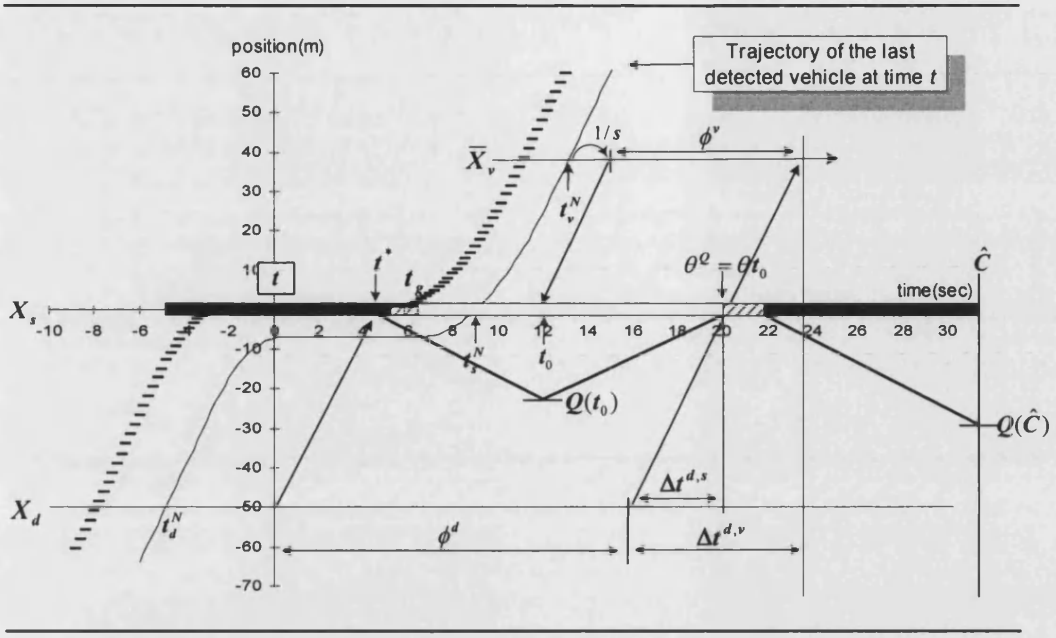


Figure 5.5 Queue clearance time

In order to find out the queue clearance time θ_j^Q for stream j ($1 \le j \le k$), the arrival and departure relationship should be considered first (5.8). As shown in Figure 5.5, starting from the trajectory of the last detected vehicle, the time when accumulated inflow is equal to accumulated outflow can be found with respect to q_j and s_j. The arrival and departure relationship between the X_d and $\bar{X}_v$ is

$$q_j \phi_j^d = s_j \phi_j^v \qquad (1 \le j \le k) \tag{5.8}$$

where ϕ_j^d is the duration of uniform arrival for stream j at X_d and ϕ_j^v is the duration of uniform departure for stream j at $\bar{X}_v$ until the queue clears,

thus, (5.8) becomes

$$\phi_j^d = \frac{s_j \phi_j^v}{q_j} \qquad (1 \le j \le k) \tag{5.9}$$

Based on this inter-relationship between arrivals and departures, the queue clearance time θ_j^Q can be estimated for each combination of signal timing plans. We assume that the uniform queue starts to build up at time t_j^*, that is the beginning time of the prediction period at X_s, and the first uniform

queue departs $1/s_j$ after t_{vj}^N at $\bar{X}_v$. The queue will clear when the following equality condition is satisfied:

$$t_{vj}^N + 1/s_j + \phi_j^v = t + \phi_j^d + \Delta t_j^{d,v} \qquad (1 \le j \le k) \tag{5.10}$$

where $\Delta t_j^{d,v}$ is the free-flow travel time for stream j from X_d to $\bar{X}_v$, $\Delta t_j^{d,v} = (\bar{X}_v - X_d)/v_0$,

thus, (5.9) becomes

$$\phi_j^d = (\frac{s_j}{s_j - q_j})(t_{vj}^N + 1/s_j - t - \Delta t_j^{d,v}) \qquad (1 \le j \le k) \tag{5.11}$$

If the calculation starts approaching from the detector position, then the queue clear time θ_j^Q is given by

$$\begin{aligned} \theta_j^Q &= t + \phi_j^d + \Delta t_j^{d,s} \\ &= t + (\frac{s_j}{s_j - q_j})(t_{vj}^N + 1/s_j - t - \Delta t_j^{d,v}) + \Delta t_j^{d,s} \end{aligned} \tag{5.12}$$

where $\Delta t_j^{d,s}$ is the free-flow travel time for stream j from X_d to X_s, $\Delta t_j^{d,s} = (\bar{X}_v - X_s)/v_0$.

As shown in Figure 5.5, first uniform arrival reaches the stop-line at time t^*. From that time, the uniform queue will start to build up vertically until its earliest departure time t_0 has been reached. If the signal is currently green and the last detected vehicle is undelayed, then $t_s^N \le t^*$, but in most other cases $t_s^N > t^*$. Hence, the uniform queue would start to build up before the last detected vehicle crosses the stop-line. The earliest departure time of the uniform queue for a stream j $(1 \le j \le k)$ at X_s can be given by

$$t_{0j} = t_{vj}^N + 1/s - \frac{\bar{X}_v - X_s}{v_0} \tag{5.13}$$

In the following section, 28 different cases of the prediction period delay are considered for a junction that has m stages in the cycle. They are formulated numerically in relation to the following components: a set of the green indications $\mathbf{t}$, a set of the red indications $\mathbf{t}_r$, a beginning time of the

prediction period t^* and an earliest time of the uniform queue dissipation begin t_0 in the interval of the lookahahead period $[t, \hat{C}]$.

5.4.2.2 Prediction period delay for the stream currently on green

There are twenty different cases of delay to be considered for a stream that is currently on green. As seen in Figure 5.3, the optimisation starts at the end of the minimum green time $t = t_1 + g_1^{\min}$ for the current stage. At time t , the indication of the signal timing sequences are always in the order of (t, t_{r1}, t_{m+1}, $\hat{C}$), where m is the maximum number of stages in the cycle, in which t_j^* and t_{0j} vary in the interval of the lookahead period $[t, \hat{C}]$. Hence, the delay of each case can be calculated first by identifying the times of t_j^* and t_{0j} in the interval $[t, \hat{C}]$, and then by estimating the cumulative queue $Q_j(t_{0j})$, $Q_j(t_{r1})$, $Q_j(t_{m+1})$ and its dissipation time $\theta_j t_{0j}$, $\theta_j t_{r1}$, $\theta_j t_{m+1}$ respectively.

The first uniform arrival of traffic begins at time t^*, and continues until the end of the lookahead period $\hat{C}$. Thus, the interval $[t_j^*, \hat{C}]$ denotes the *delay zone* for the prediction period. During that interval, the cumulative queue may all dissipate before the $\hat{C}$ has reached; if not, any residual queue $Q_j(\hat{C})$ will be included in the terminal cost function. Twenty different cases of delay when the stream is currently green can be calculated as follows:

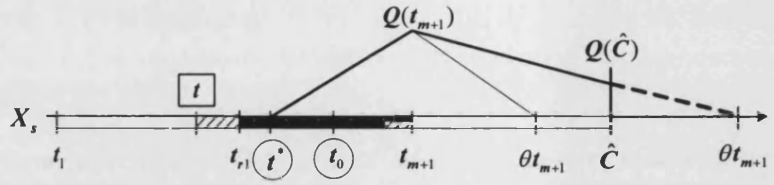
- If ($t_{r1} < t_j^* \le t_{m+1}$), there are five different cases of delay d_j^P to be considered for the stream j is currently on green, in which

  1) If ($t_j^* \le t_{0j} \le t_{m+1}$ and $\theta_j t_{m+1} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[\theta_j(t_{m+1}) - t_j^*]}{2} \tag{5.14}$$

  2) If ($t_j^* \le t_{0j} \le t_{m+1}$ and $\theta_j t_{m+1} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[t_{m+1} - t_j^*] + [\hat{C} - t_{m+1}][Q_j(t_{m+1}) + Q_j(\hat{C})]}{2} \tag{5.15}$$



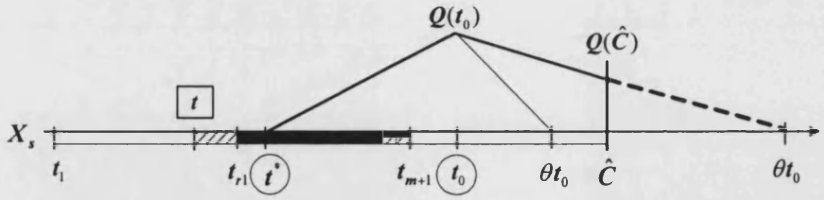
In (5.14) and (5.15), we have $Q_j(t_{m+1}) = q_j(t_{m+1} - t_j^*)$ and $\theta_j t_{m+1} = Q_j(t_{m+1})/(s_j - q_j)$, in which if $\theta_j t_{m+1} > \hat{C}$, then $Q_j(\hat{C}) = Q_j(t_{m+1}) - [(s_j - q_j)(\hat{C} - t_{m+1})]$; these values are calculated based on the vertical queueing concept (see Section 4.3). In the following calculations, the remainder will be calculated in the same way.

3) If ($t_{m+1} < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[\theta_j t_{0j} - t_j^*]}{2} \tag{5.16}$$

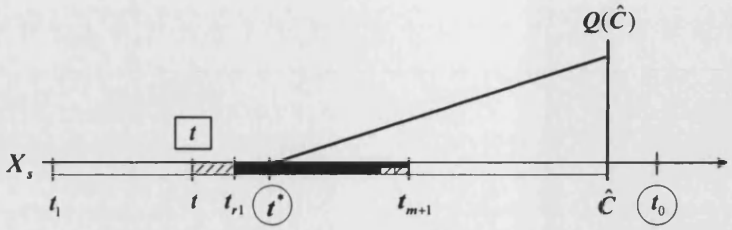
4) If ($t_{m+1} < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[t_{0j} - t_j^*] + [\hat{C} - t_{0j}][Q_j(t_{0j}) + Q_j(\hat{C})]}{2} \tag{5.17}$$



5) If ($t_{0j} > \hat{C}$), then

$$d_j^P = \frac{Q_j(\hat{C})[\hat{C} - t_j^*]}{2} \tag{5.18}$$



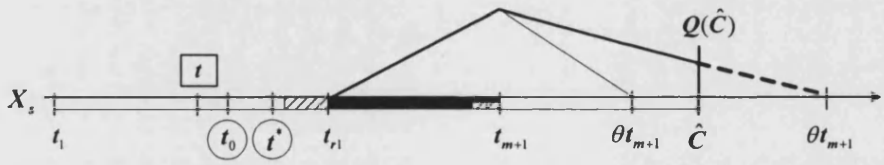
- If ($t < t_j^* \le t_{r1}$), there are eleven different cases of delay d_j^P to be considered when the stream j is currently on green, in which

1) If ($t_{0j} \le t_j^*$ and $\theta_j t_{m+1} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[\theta_j t_{m+1} - t_{r1}]}{2} \tag{5.19}$$

2) If ($t_{0j} \le t_j^*$ and $\theta_j t_{m+1} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[t_{m+1} - t_{r1}] + [\hat{C} - t_{m+1}][Q_j(t_{m+1}) + Q_j(\hat{C})]}{2} \tag{5.20}$$

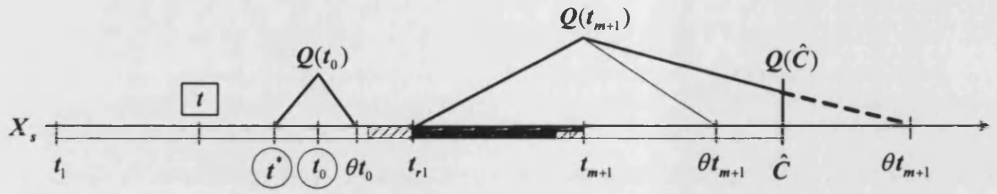


3) If ($t_j^* < t_{0j} \le t_{r1}$, $\theta_j t_{0j} \le t_{r1}$ and $\theta_j t_{m+1} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[\theta_j t_{0j} - t_j^*] + Q_j(t_{m+1})[\theta_j t_{m+1} - t_{r1}]}{2} \tag{5.21}$$

4) If ($t_j^* < t_{0j} \le t_{r1}$, $\theta_j t_{0j} \le t_{r1}$ and $\theta_j t_{m+1} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[\theta_j t_{0j} - t_j^*] + Q_j(t_{m+1})[t_{m+1} - t_{r1}] + [\hat{C} - t_{m+1}][Q_j(t_{m+1}) + Q_j(\hat{C})]}{2} \quad (5.22)$$

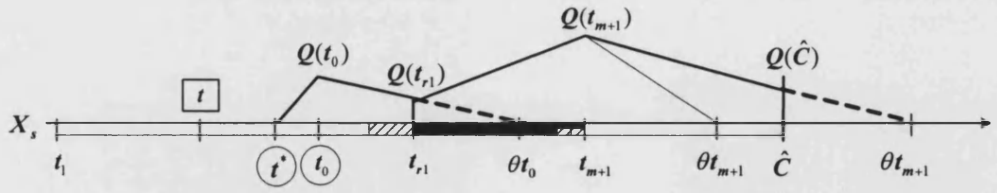


5) If ($t_j^* < t_{0j} \le t_{r1}$, $\theta_j t_{0j} > t_{r1}$ and $\theta_j t_{m+1} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[t_{0j} - t_j^*] + [t_{r1} - t_{0j}][Q_j(t_{0j}) + Q_j(t_{r1})]}{2} + \frac{[t_{m+1} - t_{r1}][Q_j(t_{r1}) + Q_j(t_{m+1})] + Q_j(t_{m+1})[\theta_j t_{m+1} - t_{m+1}]}{2} \quad (5.23)$$

6) If ($t_j^* < t_{0j} \le t_{r1}$, $\theta_j t_{0j} > t_{r1}$ and $\theta_j t_{m+1} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{0j})[t_{0j} - t_j^*] + [t_{r1} - t_{0j}][Q_j(t_{0j}) + Q_j(t_{r1})]}{2} + \frac{[t_{m+1} - t_{r1}][Q_j(t_{r1}) + Q_j(t_{m+1})] + [\hat{C} - t_{m+1}][Q_j(t_{m+1}) + Q_j(\hat{C})]}{2} \quad (5.24)$$

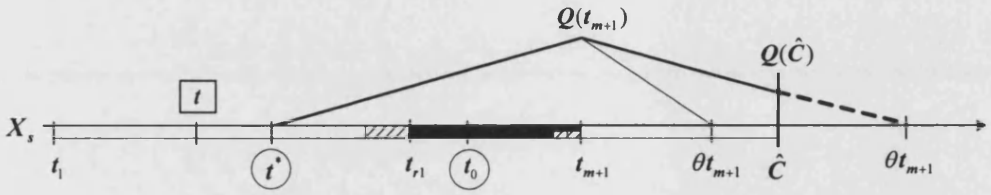


7) If ($t_{r1} < t_{0j} \le t_{m+1}$ and $\theta_j t_{m+1} \le \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[\theta_j t_{m+1} - t_j^*]}{2} \tag{5.25}$$

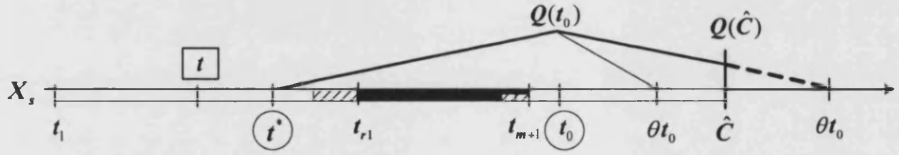
8) If ($t_{r1} < t_{0j} \le t_{m+1}$ and $\theta_j t_{m+1} > \hat{C}$), then

$$d_j^P = \frac{Q_j(t_{m+1})[t_{m+1} - t_j^*] + [\hat{C} - t_{m+1}][Q_j(t_{m+1}) + Q_j(\hat{C})]}{2} \tag{5.26}$$

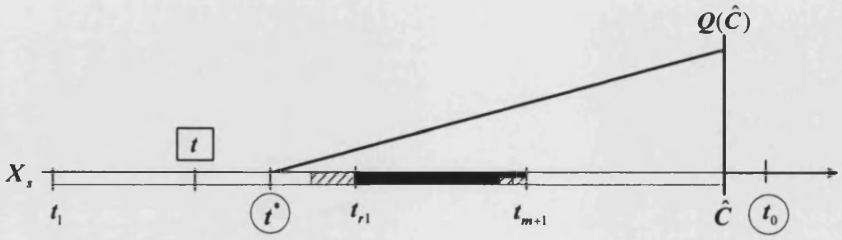


9) If ($t_{m+1} < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} \le \hat{C}$), then d_j^P is calculated by using (5.16)

10) If ($t_{m+1} < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} > \hat{C}$), then d_j^P is calculated by using (5.17)



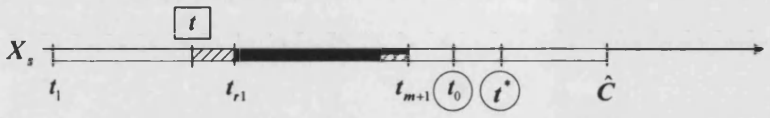
11) If ($t_{0j} > \hat{C}$), then d_j^P is calculated by using (5.18)



- If ($t_{m+1} < t_j^* \le \hat{C}$), there are four different cases of delay d_j^P to be considered when the stream j is currently on green, in which

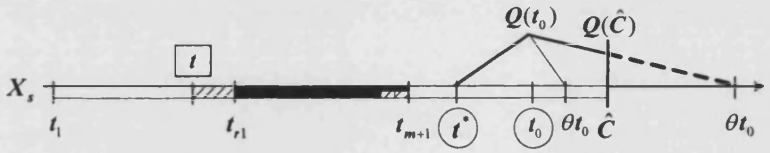
1) If ($t_{0j} \le t_j^*$), then

$$d_j^P = 0 \tag{5.27}$$

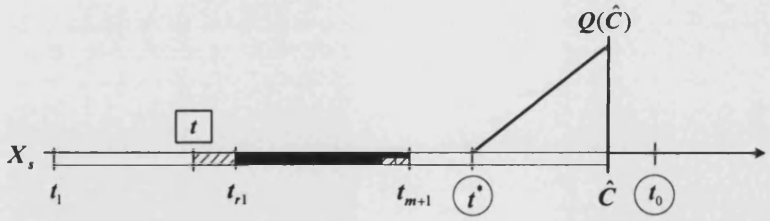


2) If ($t_j^* < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} \le \hat{C}$), then d_j^P is calculated by using (5.16)

3) If ($t_j^* < t_{0j} \le \hat{C}$ and $\theta_j t_{0j} > \hat{C}$), then d_j^P is calculated by using (5.17)



4) If ($t_{0j} > \hat{C}$), then d_j^P is calculated by using (5.18)



Finally, the total prediction period delay for those streams currently on green is given by

$$D_2^g = \sum_{j=1}^{k} d_j^P \tag{5.28}$$

and respectively the total residual queue at the end of the lookahead period $\hat{C}$ is given by

$$\hat{Q}^g = \sum_{j=1}^{k} Q_j(\hat{C}) \tag{5.29}$$

5.4.2.3 Prediction period delay for the stream currently on red

There are eight different cases of delay to be considered for a stream that is currently on red. The optimisation starts at time t. Then the indication of the signal timing sequences are in the order of (t, t_i, t_{ri}, $\hat{C}$), where $2 \le i \le m$, in which t_j^* and t_{0j} vary in the interval of the lookahead period $[t, \hat{C}]$. Hence, the delay of each case can be calculated first by identifying the times of t_j^* and t_{0j} in the interval $[t, \hat{C}]$, and then by estimating the cumulative $Q_j(t_{0j})$, $Q_j(t_{ri})$ and their dissipation time $\theta_j t_{0j}$, $\theta_j t_{ri}$ respectively. During the interval $[t_j^*, \hat{C}]$, the cumulative queue may all dissipate before the next green ends, but during the next red a new queue grows again. Thus, there are always residual queues $Q_j(\hat{C})$, which exist at the end of the $\hat{C}$. They will be included in the terminal cost function. Eight different cases of delay for a stream is currently on red can be calculated as follows:

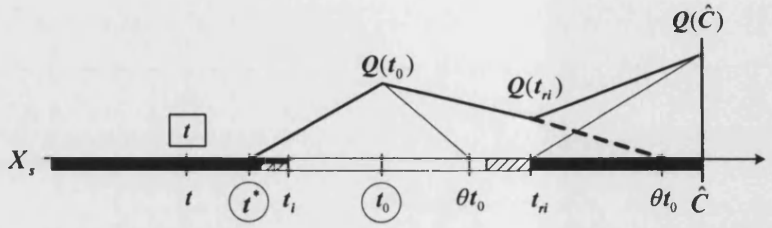
- If ($t < t_j^* \le t_i$), there are three different cases of delay d_j^P to be considered when the stream j is currently on red, in which

  1) If ($t_i < t_{0j} \le t_{ri}$ and $\theta_j t_{0j} \le t_{ri}$), then

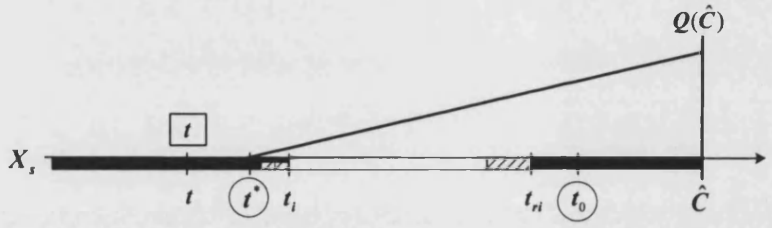
$$d_j^P = \frac{Q_j(t_{0j})[\theta_j t_{0j} - t_j^*] + Q_j(\hat{C})[\hat{C} - t_{ri}]}{2} \tag{5.30}$$

  2) If ($t_i < t_{0j} \le t_{ri}$ and $\theta_j t_{0j} > t_{ri}$), then

$$d_j^P = \frac{Q_j(t_{0j})[t_{0j} - t_j^*] + [t_{ri} - t_{0j}][Q_j(t_{0j}) + Q_j(t_{ri})] + [\hat{C} - t_{ri}][Q_j(t_{ri}) + Q_j(\hat{C})]}{2} \tag{5.31}$$



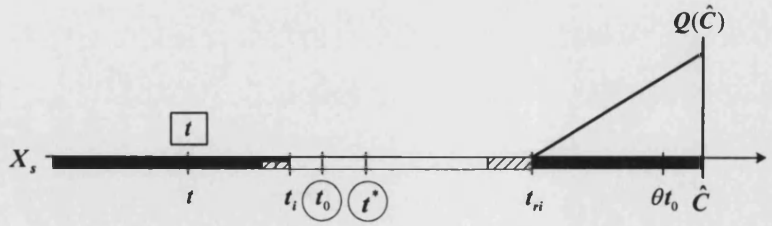
3) If ($t_{0_j} > t_{ri}$), then d_j^P is calculated by using (5.18)



- If ($t_i < t_j^* \le t_{ri}$), there are four different cases of delay to be considered when the stream j is currently on red, in which

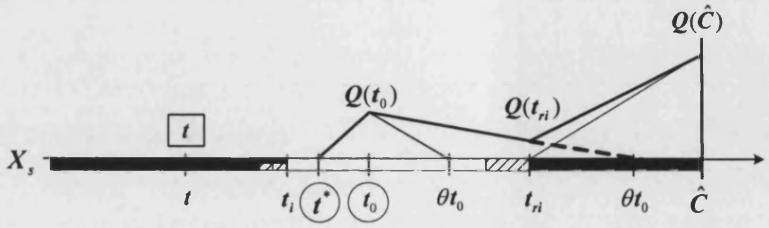
1) If ($t_i < t_{0_j} \le t_j^*$), then

$$d_j^P = \frac{Q_j(\hat{C})[\hat{C} - t_j^*]}{2} \tag{5.32}$$

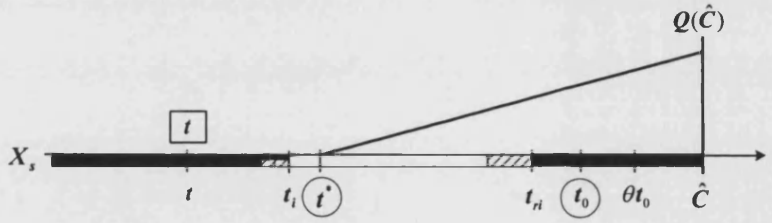


2) If ($t_j^* < t_{0j} \le t_{ri}$ and $\theta_j t_{0j} \le t_{ri}$), then d_j^P is calculated by using (5.30)

3) If ($t_j^* < t_{0j} \le t_{ri}$ and $\theta_j t_{0j} > t_{ri}$), then d_j^P is calculated by using (5.31)

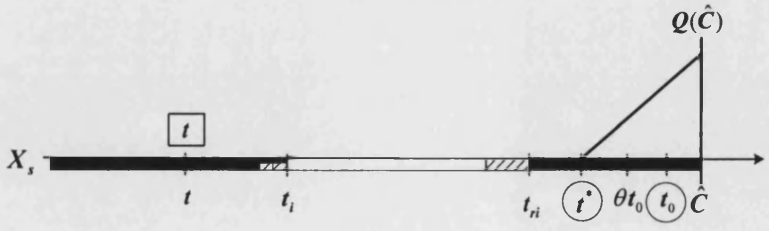


4) If ($t_i < t_{0j} \le \hat{C}$), then d_j^P is calculated by using (5.32)



- If ($t_{ri} < t_j^* \le \hat{C}$), there is one case of delay to be considered when the stream j is currently on red, in which

1) If ($t_j^* < t_{0j}$), then d_j^P is calculated by using (5.18)



Finally, the total prediction period delay for those streams currently on red is given by

$$D_2^r = \sum_{j=1}^{k} d_j^p \tag{5.33}$$

and respectively the total residual queue at the end of the lookahead period $\hat{C}$ is given by

$$\hat{Q}^r = \sum_{j=1}^{k} Q_j(\hat{C}) \tag{5.34}$$

5.4.3 Terminal cost function

This is an estimate of the additional delay caused by residual queues at the end of the lookahead period $\hat{C}$, which consists of the residual vehicles $\bar{N}_0^g$ (5.6) and $\bar{N}_0^r$ (5.7) from the detection period, and additional built-up queues $\hat{Q}^g$ (5.29) and $\hat{Q}^r$ (5.34) from the prediction period. In practice, there will almost certainly be some residual queues at the end of $\hat{C}$. Thus, this is represented in the terminal cost function. A simple form of a terminal cost function developed by Robertson and Bretherton (1974) is adopted for use here. This is an empirically approximated quadratic function, which estimates the additional delay caused by the queues at the end of the lookahead period. The terminal cost function of Robertson and Bretherton is established as

$$TC = \frac{0.2}{1-Y}(Q_g + 1.3Q_r)^2 \tag{5.35}$$

where

TC is the terminal cost caused by the initial queue value greater than zero and it has units of vehicle-intervals (Robertson and Bretherton used 5 seconds per interval),

Q_g is the sum of residual queues for streams which will receive a green indication at the end of the lookahead period ($Q_g = \bar{N}_0^r + \hat{Q}^r$),

Q_r is the sum of residual queues for streams which will receive a red indication at the end of the lookahead period ($Q_r = \bar{N}_0^g + \hat{Q}^g$),

Y is the sum over representative streams of the proportion of the arrival flows to their saturation flow rates ($Y = \sum_{j=1}^{k} \frac{q_j}{s_j}$).

5.4.4 Objective function (Estimates of control performance)

Finally, we have the total delay incurred across all streams at the junction. The three distinct components of delay are: the detection period delay D_1 (5.5), the prediction period delay D_2^g (5.28), and D_2^r (5.33) and the terminal cost TC (5.35). However, these delays have units of vehicle-seconds, so that they need to be converted to rate of delay by dividing their final value by the duration of the lookahead period about one-cycle $\phi\hat{C}$. As discussed in Section 2.2, the rate of delay is the excess number of vehicles in the vicinity of the junction due to the operation of signals. It is normally expressed as the difference between the average number of vehicles within the region of influence of the junction, and the number that would be there if the same flow of traffic were unimpeded by the signals. This is the average delay incurred per unit time by traffic on the stream. It has units of vehicles, thus the rate of delay is additive over the streams.

By summing the delay obtained from equations (5.5), (5.28), (5.33) and (5.35), the total rate of delay caused during the lookahead period of about one-cycle $\phi\hat{C}$ (5.4) with respect to the signal timing plan $\mathbf{H}$ is given by

$$D(\mathbf{H}, \phi\hat{C}) = \frac{D_1 + D_2^g + D_2^r + TC}{\phi\hat{C}} \tag{5.36}$$

The objective function can be set at the choice of users. Three different possibilities for the objective function that are considered here are:

- Detection period delay: $D(\mathbf{H}, \phi\hat{C}) = D_1 / \phi\hat{C}$,
- Detection period delay plus Prediction period delay: $D(\mathbf{H}, \phi\hat{C}) = (D_1 + D_2^g + D_2^r) / \phi\hat{C}$,
- Detection period delay plus Prediction period delay plus Terminal Cost: $D(\mathbf{H}, \phi\hat{C}) = (D_1 + D_2^g + D_2^r + TC) / \phi\hat{C}$.

We note that in each of these cases, the objective function is non-linear. Because the optimisation method that is developed here performs a direct search over values of the objective, it can accommodate this kind of formulation.

5.5 DISCUSSION

In this chapter, a systematic optimising method for us in dynamic signal control was presented. The optimising method developed here can be defined as a stage-based exhaustive search method in the rolling horizon concept. The optimising method consists of two parts: the dynamic optimiser and the traffic model. During the optimisation process, the control decisions are made on the basis of an exhaustive search method, in which the dynamic optimiser provides different combinations of signal timing plans for the near future to the traffic model, and the traffic model provides estimates of operational performance back to the dynamic optimiser. Finally, the dynamic optimiser makes control decisions whether or not to extend the current stage. The key feature in this approach is that a plan is developed for the entire lookahead period, but only the first part (one time-step) is implemented. The lookahead period proposed here is the same for all streams, but varies according to the state of the dynamic optimiser. The lookahead period represents a compromise between the need to consider the future consequences of control decisions implemented in the short term and the limited availability of detailed data concerning vehicular arrivals. Moreover, this kind of search method will identify the best plan at that time for the objective in hand, though further arrivals may render that sub-optimal in the future.

Various traffic models can be used with this optimiser, and two have been investigated here (KCS traffic model and Vertical qeueing model). Furthermore, various combinations of three distinct components of the objective function (detection period delay, prediction period delay and terminal cost) can be used. The optimisation formulation for traffic-responsive signal control presented here makes use of features provided by a number of different approaches to dynamic signal optimisation. The formulation is developed on the stream-by-stream basis, so that it is applicable to a wide range of junction configurations.

As noted, the exhaustive search can in some cases be time consuming because it searches all possible extensions of the current stage in combination with the durations of all stages in the sequence. At each time-step, the optimiser estimates the mean rate of delay over the lookahead. Sometimes and very rarely, the decision trees could be too large to be searched completely within

the available time period. As an alternative to the exhaustive search method, we can suggest a heuristic search method. At time of optimisation, the extension test would be done only for the current stage, during which the end of the lookahead period is estimated based on the queue clearance time of each following stage. Thus, we can take a lookahead period that is long enough to serve all queueing vehicles in whole stages. This kind of heuristic method would be simpler and less time consuming in optimising process, but would have the possibility of giving sub-optimal decisions in some cases.

The optimising method presented here will be tested in the Chapter 6 by interfacing it with the SIGSIM microscopic traffic simulator, in which the performance of the proposed method will be investigated and compared with existing control methods.

6. PERFORMANCE ANALYSIS

6.1 INTRODUCT ION

In this chapter, we compared the performance of the dynamic optimising method that was presented in Chapter 5 with existing methods (System D Vehicle Actuated and fixed-time control). In order to undertake this comparison, the novel optimising method was interfaced with SIGSIM microscopic traffic simulator as a new control strategy.

6.2 METHODOLOGY

Eight different control strategies were tested for fourteen different combinations of flow patterns (7 balanced and 7 unbalanced) by using the SIGSIM traffic simulator. Each control strategy was run for 10 times with a same initial random number seed (use of the same seed number generates a same number of vehicles). The final results were presented in terms of the mean rate of delay over all streams at the junction with associated standard error for 10 runs. Finally, the estimates of performance for different control strategies were analysed statistically, based on the two sample means $t-$test and the paired mean $t-$test. For this examination, six different combinations of optimising methods are considered: a dynamic optimiser with two traffic models (KCS traffic model and Vertical queueing model) and three different combinations of the objective function (Detection period delay, Prediction period delay and Terminal Cost). The comparison scenarios are depicted in Figure 6.1.

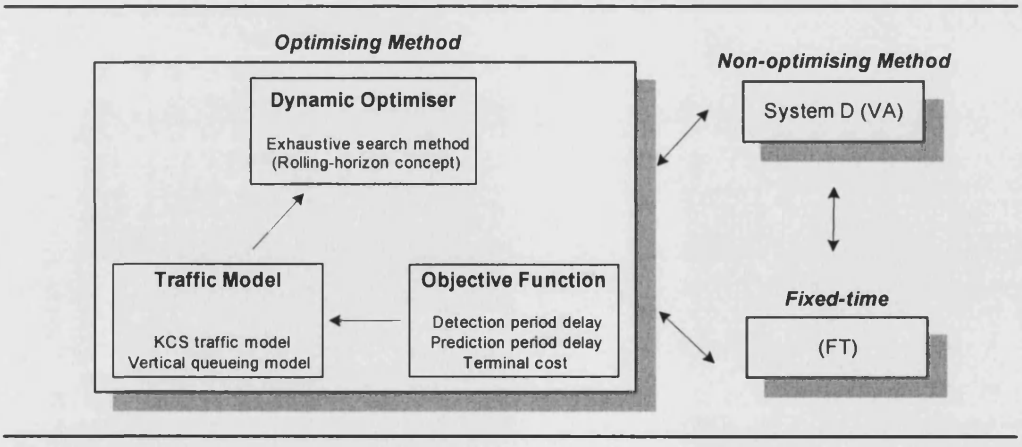


Figure 6.1 Comparison scenarios

Before running the optimising method in the SIGSIM, we need to estimate the saturation flow from the pilot simulation with a set of predefined input data. As we noted, SIGSIM does not need the saturation flow value, but it is influenced by the user-defined speed reduction factor (see Section 3.6.3). As part of the simulation results, SIGSIM provides estimates of the saturation flow, the start lag and the end lag for each lane according to the Road Note 34 (RRL, 1963) method. Based on these estimates, we can obtain the mean saturation departure time and acceleration rate.

The reason for deriving these values from the pilot simulation is that some parameters are required by traffic models to represent the motion of vehicles. Hence, we can minimise the discrepancy of traffic flow between the numerical estimation and the simulation. By adopting these representative parameters into the traffic model, we can suppose that the interfaced optimising method has a similar traffic flow condition to the SIGSIM traffic simulator. As presented in Chapter 4, the KCS traffic model requires an acceleration rate and braking rate, while in contrast the Vertical queueing model requires a start lag.

6.3 SIMULATION DESIGN

6.3.1 Input data in SIGSIM

The example calculations were performed for a single junction that has 8 links, and each link has 1 lane. As shown in Figure 6.1, the junction comprises 4 entry links and 2-stage signal control. Each of these entry links has traffic flows in one direction, and all traffic goes straight ahead at the junction. In the simulation, 4 entry links (102, 104, 106 and 108) are controlled by different signal control strategies, but the estimates of performance are evaluated for all 8 links at the junction.

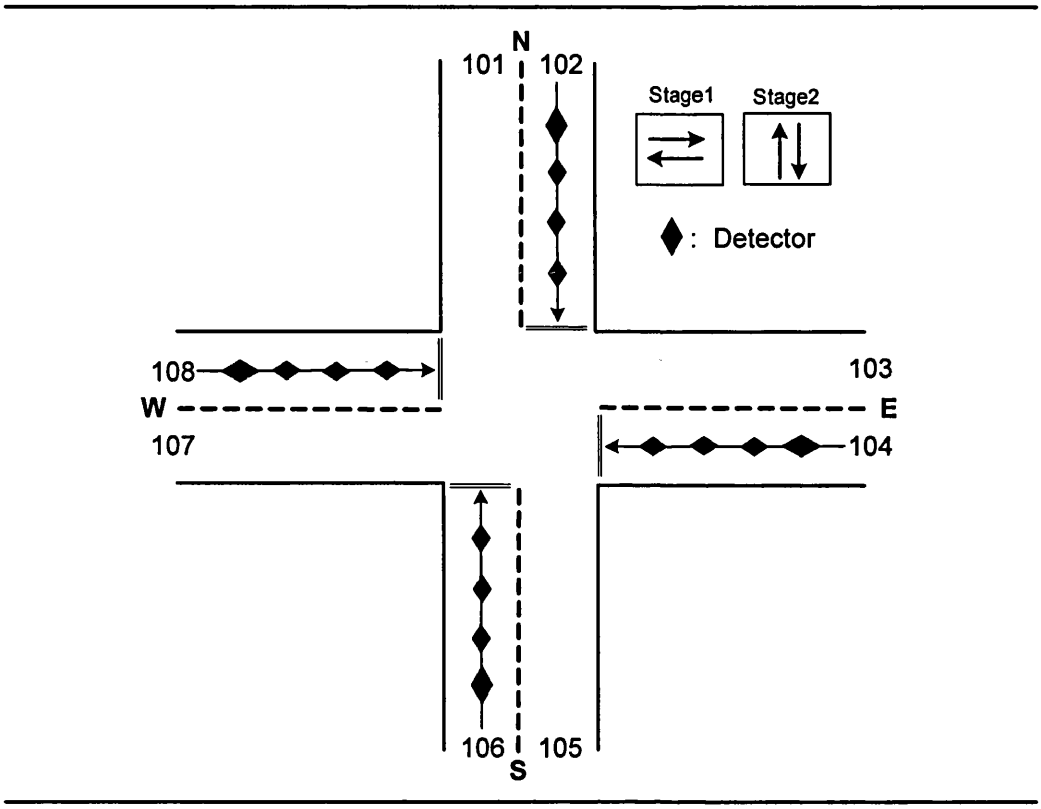


Figure 6.2 Junction layout for the simulation

SIGSIM has five input data files. Among them, <junction.file> specifies link length, detector specification, speed reduction, vehicle profile, duration of simulation, random number seed and number of runs. <signal.file> specifies controller scan time, signal control policy, duration of green (minimum and maximum green), phases in stages and interstage time. Other parameters (i.e. desired speed, vehicle length, maximum acceleration rate, maximum braking rate, etc) can be specified in the Header file <symtype.h>. For this examination, the input data specified in the SIGSIM are as flows:

1) Link length and detector specification:

- Vehicles are generated at the end of each link, which is set to 400 m.
- For VA control: three detectors are placed at 12 m, 26 m and 40 m.
- For the optimising method (KCS traffic model and Vertical queueing model); one detector is placed at 50 m.
- Length of detector is 2 m.

2) Speed reduction (saturation flow):

- According to the Road Note 34 (RRL, 1963) and Kimber et al (1986), for a lane 3.25m wide with no turning traffic and with 0% gradient, the standard saturation flow is approximately 2080 vehicles/h.
- Based on cars travelling with 15m/s, that saturation flow can be maintained in SIGSIM by introducing the speed reduction 25% (see Section 3.6.3 and 6.2.2.1 for detailed discussion).

3) Vehicle profile: One type of passenger car is used in this examination.

Vehicle type	Length (m)	Desired speed (m/s)	Maximum acceleration (m/s^2)	Maximum braking (m/s^2)
Car	6.5 (0.3)	15.0 (3.2)	1.7 (0.3)	3.4 (0.3)

Remarks) Figures in brackets represent standard deviations.

4) Duration of simulation, random number seed and number of runs:

- Simulation time is set to 4200 seconds, within which the warm-up time is 600 seconds.
- The initial random number seed is set to 10. This controls vehicle generation with respect to input flow data (the same seed number with the same input flow generates same number of vehicles).
- Number of runs is set to 10 times (the same number of runs with identical data will produce identical output).

6) Controller scan time: time-step is set to 0.5 seconds.

7) Signal control policy: 8 different control strategies are tested with 8 different inflow conditions in the range from 200 to 800 vehicles/h for each lane. Here, the objective function has a set of choices: Detection period delay (D), Detection + Prediction period delay (D+P) or Detection + Prediction + Terminal Cost (D+P+TC).

- Case 1 (Fixed time control): Optimised fixed-time control (FT).
- Case 2 (Non-optimising method): System D Vehicle Actuated (VA).
- Case 3,4,5 (Optimising method): Dynamic optimiser + KCS traffic model with one from three objective functions, which are denoted as KCS (D), KCS (D+P) and KCS (D+P+TC).

- Case 6,7,8 (Optimising method): Dynamic optimiser + Vertical queueing model with one from three objective functions, which are denoted as Vertical(D), Vertical(D+P) and Vertical(D+P+TC).

8) Phases in stages:

- As seen in Figure 6.2, the experimental junction controls with two stages: stage1 (102 and 106) and stage 2 (104 and 108).
- Interstage time is set to 5 seconds (2 seconds of red and amber and 3 seconds of amber).

9) Duration of green:

- Minimum green is set to 7 seconds.
- Maximum green is set to 20 seconds.

6.3.2 Estimation of input parameters for the optimising method

As noted in Section 3.6, SIGSIM is a microscopic simulator. The speed and position of vehicles are updated according to the time-by-time simulation results. In contrast, the traffic model that will be used in the optimising method is a homogeneous numerical equation. The motion of vehicles is estimated according to the standard formulae. Thus, we need some parameters which can represent the motion of vehicles in SIGSIM, and we need to minimise the discrepancy of traffic flows between the numerical estimation and the simulation. By adopting the SIGSIM estimated parameters into the traffic model, we can suppose that the interfaced optimising method has a similar traffic flow condition to the SIGSIM traffic simulator.

The formulae established for the KCS traffic model require one detector for each lane and five motion parameters (e.g. desired speed, acceleration rate, braking rate, saturation departure time and vehicle length). In contrast, the Vertical queueing model requires one detector for each lane but three motion parameters (e.g. desired speed, start lag and saturation departure time). From the pilot simulation, SIGSIM estimates the saturation flow, the start lag and the end lag. Based on them, the saturation departure time corresponding to the saturation flow can be obtained, and as case the acceleration rate corresponding to the start lag. The end lag is used for extra green time taken from the amber. In the following estimation, we will assume the desired speed and the braking rate is same as the SIGSIM used.

6.3.2.1 Saturation flow, start lag, and end lag

In SIGSIM, a saturation flow, a start lag and an end lag for each lane are calculated according to the Road Note 34 (Road Research Laboratory, 1963) method. In this method, these three values are calculated based on the average number of vehicles that have crossed the stop-line in each of three counting intervals.

When a traffic signal turns green, vehicles take some time to start (reaction time, τ =0.67 seconds) and then accelerate to normal running speed, but after this initial period the queue discharges at a constant rate. The *saturation flow* is the flow which would be obtained if there was a continuous queue of vehicles and they were given 100 per cent green time: it represents the maximum number of vehicles that can be passed in one hour, assuming a continuous display of green and a continuous queue of vehicles. In the following examination, the time from the start of red and amber until vehicles start to depart with this maximum rate is called the duration of the *start lag*. The time taken from the discharging vehicles during the amber period is called the duration of the *end lag*. In general, these values can be estimated either by field measurement or by calculation using empirical formulae. The field measurement demands a high number of observations, but it estimates the saturation flow accurately. In contrast, the empirical formula is easier to use, but it does not always lead to an accurate estimation, because it is based on averages over a range of cases, and may not be valid for a particular case.

1) The Road Note 34 method in SIGSIM

According to the Road Note 34 method, saturation flows are calculated only cases of a fully saturated green period (i.e. a green period during which the queue is not fully discharged). The observations should extend over at least 30 cycles, of which there should be a sufficient number of fully saturated cases (about 15). If no cycles are saturated for a lane, then SIGSIM does not calculate saturation flow or lags. By investigating the source code <Rn34.c>, it became clear that the use of the formulae is correct, but there was a mistake in the way that the vehicles were recorded. The original source code provides relatively high saturation flow, and thus it underestimates the start lag and the end lag. In this analysis, the amended source code will be used for recording vehicles crossing the stop-line in each of three intervals are as follows:

- As shown in Figure 6.3, the first interval begins from the start of red and amber T_0, from which the vehicles n_1 that passing through the stop-line during the first 10 seconds ($T_1 - T_0$) are recoded for the start lag estimation. The duration of the first interval is ϕ_1. Road Note 34 recommended 6

seconds for the first interval, but we took 10 seconds because for free flow speed 15 m/sec and acceleration rate 1.7 m/sec^2, it takes about 8.8 seconds of acceleration time until the foremost vehicle in the queue regains free flow speed.

- The second interval begins from the first vehicle passing the stop-line after T_1. In this interval, the vehicles n_2 that passing through the stop-line until the start of the amber T_2 are recorded for the saturation flow rate (vehicles/sec) estimation. The duration of the second interval is ϕ_2.

- Vehicles n_3 passing the stop-line during the amber period and early in the red indication period are recorded for the end lag estimation. The duration of the third interval is ϕ_3.

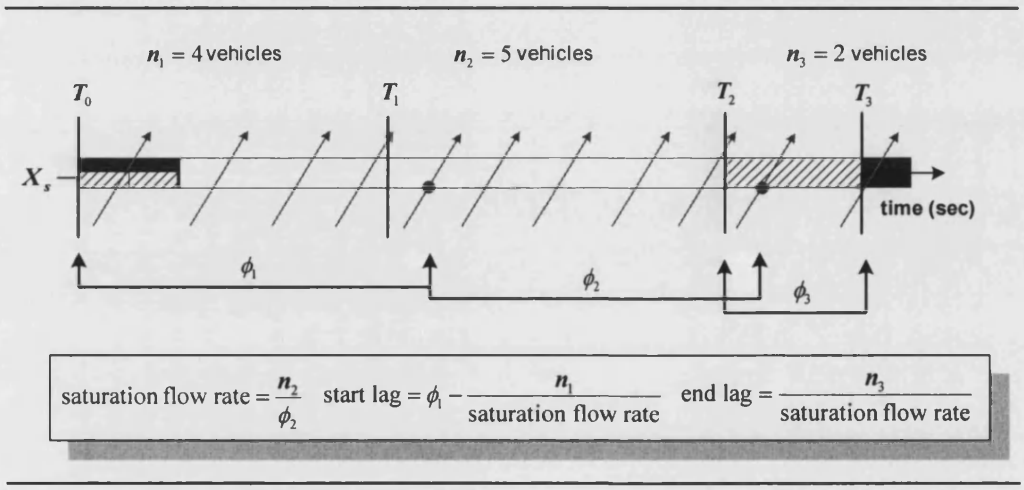


Figure 6.3 Saturation flow rate, start lag and end lag estimation in the Road Note 34 method

2) Speed reduction against saturation flow in SIGSIM

Based on the input data presented in Section 6.2.1 (e.g. desired speed of cars 15 m/s, acceleration rate 1.7 m/s^2 and braking rate 3.4 m/s^2), one lane with straight movement was tested for the maximum green of 20 seconds and 30 seconds. One hour's traffic was simulated by increasing the speed reduction in the range of 0% to 90%. As was discussed in Section 3.6.3, this speed reduction applies to each vehicle within a specific zone of influence and it controls saturation flows. As shown in Figure 6.4, SIGSIM produces saturation flows around 2000 vehicles/h in the range of 20 to 40% of the speed reduction and none of the results is higher than 2500 vehicles/h.

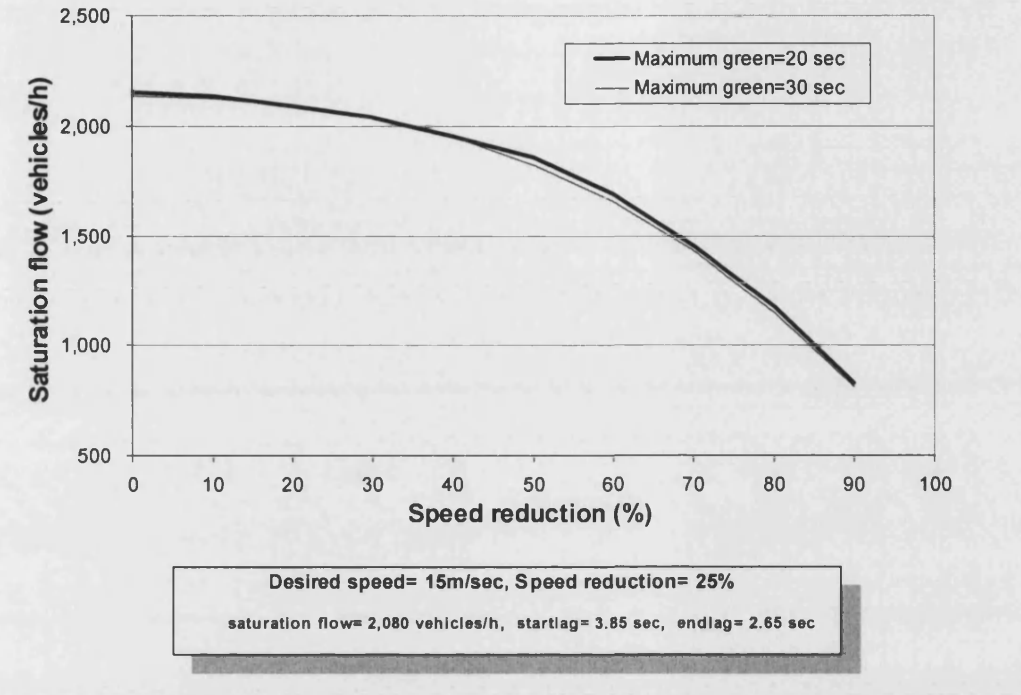


Figure 6.4 Speed reduction against saturation flow in SIGSIM

3) Pilot simulation results (saturation flow, start lag and end lag)

The saturation flow for a lane 3.25 m wide in a good condition is 2080 vehicles/h (Kimber et al, 1986). From these pilot simulation results, the speed reduction of 25% produces the saturation flow of about 2080 vehicles/h with a start lag of 3.85 seconds and an end lag of 2.65 seconds.

6.3.2.2 Acceleration rate

According to the given input parameters in SIGSIM, such as the desired speed v_0, reaction time τ and maximum acceleration rate a, we can get the start lag $\hat{l}_s$ from the pilot simulation. By fixing the desired speed, the reaction time and the start lag, we can calculate the mean acceleration rate $\hat{a}$ that will be used in the KCS traffic model. As shown in Figure 6.5, the position for regaining the free-flow speed is $\bar{X}_v = v_0^2 / 2\hat{a}$ and its duration is $v_0 / \hat{a}$ (equations 4.6 and 4.3). The relationship between the start lag and the acceleration rate in the KCS traffic model is

$$\hat{l}_s = \tau + \frac{v_0}{2\hat{a}} \tag{6.1}$$

Using (6.1), the acceleration rate $\hat{a}$ can be calculated by

$$\hat{a} = \frac{v_0}{2(\hat{l}_s - \tau)} \tag{6.2}$$

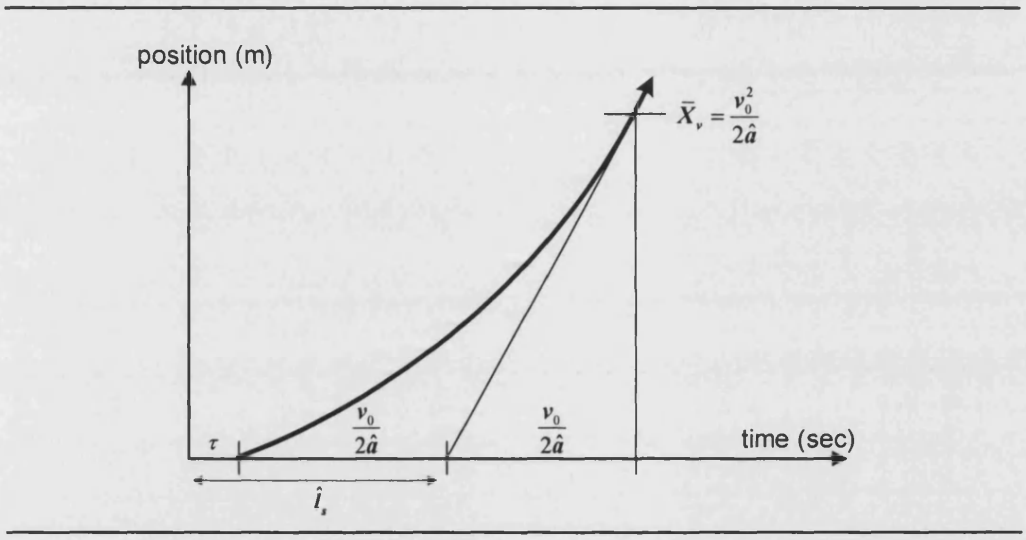


Figure 6.5 Calculation of acceleration rate using the estimated start lag

6.3.2.3 Input parameters in the KCS traffic model and the Vertical queueing model

Using the input data specified in Section 6.2.1 and the speed reduction of 25%, the pilot simulation produced the saturation flow about 2080 vehicles/h, the start lag of 3.85 seconds and the end lag of 2.65 seconds. Based on this, we could calculate the mean acceleration rate, which is 2.36 m/s^2 (6.2). The derivation of input parameters for the KCS traffic model and the Vertical queueing model are estimated as follows:

- According to the given desired speed 15m/s and the acceleration rate 1.7 m/s^2, the expected start lag from the beginning of the red and amber is about 5 seconds (6.1), including the reaction time τ = 0.67 seconds. In general, this lag represents the lost time incurred by vehicles in the queue.

- However, the SIGSIM simulation produced a start lag of about 3.85 seconds. This value is smaller than we expected. According to the Road Notes 34 method, start lag is estimated on the basis of the saturation flow. This lag represents the lost time incurred during the red and amber plus the whole green periods.

- As discussed from the beginning of Section 6.2.2, the traffic models that will be used in this examination have homogeneous numerical equations. Thus, some necessary adjustment is unavoidable in order to minimise the discrepancy between the numerical estimation and the simulation. As we noted, the KCS traffic model controls saturation flow at the downstream free-flow position $\overline{X}_v$, but the Vertical queueing model controls the saturation flow at the stop-line position X_s. For the Vertical queueing model, we do not need to make any parameter adjustments. In contrast, we need a new acceleration rate for the KCS traffic model.

- Based on the start lag estimated from the SIGSIM simulation, we could calculate the mean acceleration rate that can match the saturation flow condition in SIGSIM. For this examination, the acceleration rate that will be used in the KCS traffic model is 2.36 m/s^2 (6.2), which is bigger than the SIGSIM traffic simulator uses (1.7 m/s^2).

- Vehicle length used is 7m, which includes a safety margin of 0.5m.

- Traffic models do not directly take the end lag in the trajectory formulae, but this value is adopted by the optimising method to allow some vehicles during the amber time period. The end lag adopted is 2.65 seconds.

The input parameters that will be used in the KCS traffic model and the Vertical queueing model are shown in Table 6.1.

Table 6.1 Input parameters in the KCS traffic model and the Vertical queueing model

Traffic model	Vehicle Length (m)	Desired speed (m/s)	Acceleration rate (m/s^2)	Braking rate (m/s^2)	Startlag (sec)	Endlag (sec)	Saturation Departure time (sec/veh)
KCS	7*	15.0*	2.36^	3.4*	-	2.65^	1.73^
Vertical	-	15.0*	-	-	3.85^	2.65^	1.73^

Remarks) *: exogenous parameters, ^: estimated parameters from the pilot simulation, -: not required.

6.3.3 Investigation of vehicle motions

Finally, we have all the representative parameters that are required by the optimising methods. Before running the optimisation, we should investigate one last thing about vehicle motion: whether or not the trajectory estimated from the KCS traffic model and the Vertical queueing model conforms with the movement of vehicles in the SIGSIM traffic simulator.

According to the pilot simulations, it is clear that the numerical formulae can closely describe the motion of vehicles in the SIGSIM simulation. In the fully saturated case, an average of about 13 vehicles cross the stop-line during the 25 seconds of one stage time period, which includes 2 seconds of red and amber, 20 seconds of green and 3 seconds of stopping amber (see Table 6.2 and Figure 6.6). The motion of vehicles in the SIGSIM simulation and in the two traffic models are discussed below:

1) SIGSIM simulation

- The first vehicle is modelled as responding immediately to the change of lights. Thus, it departs about 0.73 ~ 0.90 seconds after the beginning of red and amber.
- The following vehicles cross the stop-line with a relatively short headway.

2) KCS traffic model

- The first vehicle departs 0.67 seconds after the beginning of red and amber.
- After the 7^{th} vehicle, further vehicles cross the stop-line at saturation departure time.
- After the 7^{th} vehicle, the KCS traffic model differs from SIGSIM with slightly lower values, but they are not significantly different.

3) Vertical queueing model

- The first vehicle departs 3.85 seconds after the beginning of red and amber.
- All the following vehicles cross the stop-line at saturation departure time.
- After the 6^{th} vehicle, further vehicles departure times at the stop-line are almost same as the KCS traffic model.

Table 6.2 Departure time of vehicles at the stop-line (fully saturated case)

	Crossing time at the stop-line (1st ~ 13th vehicle)												
	1	2	3	4	5	6	7	8	9	10	11	12	13
SIGSIM	0.73	4.33	6.57	8.55	10.36	12.34	13.98	15.54	17.29	18.85	20.58	22.19	23.36
KCS	0.67	4.35	6.61	8.65	10.56	12.39	14.17	15.91	17.64	19.37	21.10	22.83	24.56
Vertical	3.85	5.58	7.31	9.04	10.77	12.50	14.23	15.96	17.69	19.42	21.15	22.88	24.61

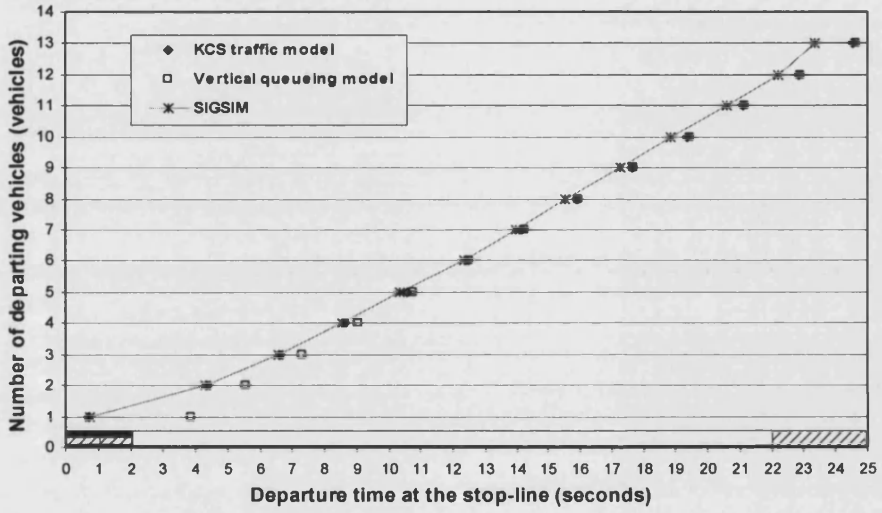


Figure 6.6 Departure time comparison at the stop-line (a fully saturated case: KCS traffic model, Vertical queueing model and SIGSIM)

6.4 METHODS FOR ESTIMATING PERFORMANCE AND VALIDATION

SIGSIM is a microscopic simulation program that can be used to simulate the traffic behaviour of individual vehicles for a signal controlled junction under different signal control strategies. One important feature of SIGSIM is that it enables a measure of 'between runs variability', which means that if SIGSIM is set to run more than once with a same input data, including identical random number seed, it will produce the same results. This feature is useful, and it is important when comparing different control strategies that a number of runs are carried out to establish the degree of variation arising from random variations in traffic flow patterns. As part of the final results, SIGSIM

provides an operational performance (mean rate of delay) that is qualified by the associated standard error to the number of runs. In this section, we discussed how SIGSIM estimates the performance and how the results can be validated.

6.4.1 How SIGSIM and the optimising method communicate

As discussed in Section 3.6.1, SIGSIM is an event-based simulation model. Thus, simulation is only updated at the occurrence of an event. According to the user defined time-step (controlscan event time), SIGSIM updates continuously the time of events, light status and detector data. During this time-step, if any new events are generated, this event is stored into the current event list and the earliest of them is identified for the next event.

For this examination, the time-step is set to 0.5 seconds. At the end of each time-step, the SIGSIM traffic simulator updates the light status of all lanes at the junction according to the control decision made from the dynamic optimiser. In each optimisation, the dynamic optimiser provides the signal timing plans to the traffic model. Meanwhile, the traffic model extracts from SIGSIM the time of vehicles crossing the detector, and provides the estimated delay back to the dynamic optimiser. In this alternating order, the dynamic optimiser makes control decisions whether or not extend the current stage and send that decision to SIGSIM. Finally, SIGSIM provides the simulation results in terms of the mean rate of delay. The interface between the SIGSIM and the optimising method is depicted in Figure 6.7.

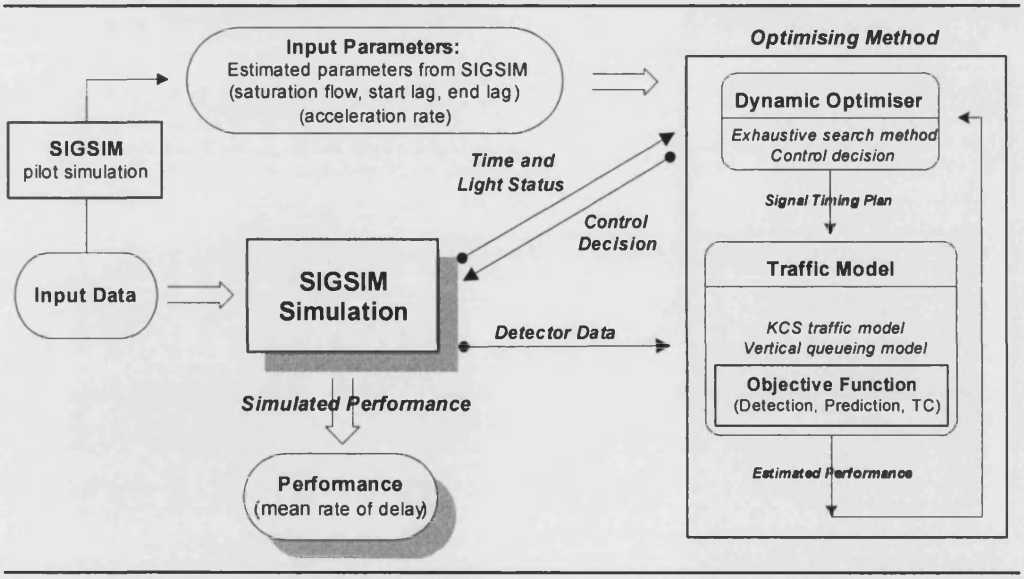


Figure 6.7 SIGSIM and optimising method interface

6.4.2 Two sample means $t-$test and the paired mean $t-$test

When we compare the difference between two control strategies, we should assess whether or not the observed difference is statistically significant. The $t-$test can be used to assess whether the difference between two means is statistically significant. In this section, we consider a practical case where the population variance is unknown. As we described earlier, each control strategy will be tested for 7 different inflow scenarios in the range from 200 to 800 vehicles/h per lane and run for 10 times. For two chosen control strategies, *two-sample means $t-$test* can compare the mean difference between same inflows and the *paired mean $t-$test* can compare overall paired mean differences.

Formulating the problem more generally, two population means are commonly compared by forming their difference ($\mu_1 - \mu_2$). A reasonable estimate of this is the corresponding difference in sample means ($\overline{X}_1 - \overline{X}_2$) and the estimated variances of S_1^2 and S_2^2, on the basis of independent random samples of size n_1 and n_2. We adopt a one-sided $t-$test because in each comparison we want to support the alternative hypothesis $\mu_1 > \mu_2$. If t-value exceeds the critical value t_α, the null hypothesis must be rejected. Thus, with 95% confidence we conclude that the average delay of the first class is greater than the second class and they are statistically significant.

1) Two sample means $t-$test

In practice, when sample size n_1, n_2 or both are small and the population variances are unknown, we can use the pooled estimate of the common variance S_p^2 by assuming that both populations are normal with $\sigma_1 = \sigma_2$. Since both populations have the same variance σ^2, it is appropriate to pool the information from both samples with $n_1 + n_2 - 2$ degrees of freedom. The formula for the two sample mean $t-$test with unknown variance is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{6.3}$$

where $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ ($S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2}{n_1-1}$ and $S_2^2 = \frac{\sum_{i=1}^{n_2}(X_i - \bar{X}_2)^2}{n_2-1}$ for observed value X_i).

The testing hypotheses and other criteria for the two sample mean $t-$test are as follows:

- Null hypothesis: $\mu_1 - \mu_2 = 0$, Alternative hypothesis: $\mu_1 - \mu_2 > 0$ (adopt one-sided $t-$test).
- Level of significance: α = 0.05 (95% confidence level).
- Criterion: Reject the null hypothesis if $t > t_\alpha$

  (i.e. t_α = 1.73 with degree of freedom $n_1 + n_2 - 2 = 18$, where $n_1 = n_2 =$ 10 runs).

2) Paired mean $t-$test

In the case where there is an association between the elements of one sample and those of another, a paired mean $t-$test can be used. The first step is to calculate the paired mean difference of each comparing group. Thus, let the pair of means $(\bar{X}_i, \bar{Y}_i)$ denote the delay for the control strategy I and the control strategy II for each inflow i =1, 2,...n. The paired statistical analysis proceeds by considering the differences $D_i = \bar{X}_i - \bar{Y}_i$ (for $i = 1, 2, ...n$).

The formula for the paired mean $t-$test is

$$t = \frac{\bar{D}}{S_D / \sqrt{n}} \tag{6.4}$$

where $\bar{D}$ is the mean of D_i, S_D is the standard deviation of D_i and n is a sample size.

The testing hypotheses and other criteria for the paired mean $t-$test are as follows:

- Null hypothesis: $\mu_D = 0$, Alternative hypothesis: $\mu_D > 0$ (adopt one-sided $t-$test).
- Level of significance: α =0.05 (95% confidence level).

- Criterion: Reject the null hypothesis if $t > t_\alpha$

  (i.e. $t_\alpha = 1.94$ with degree of freedom $n-1=6$, where $n=7$ paired samples).

6.5 SIMULATION RESULTS

Each of the test runs was evaluated over 1 hour of simulation time after 10 minutes of simulation warm-up time. They were run for ten times with inflows ranging from 200 to 800 vehicles/h in each lane. By fixing the minimum green time as 7 seconds and the maximum green time as 20 seconds, each inflow pattern had equal degree of saturation in the range from 0.3 to 1.0 (from fairly light to heavy). The test cases included both balanced and unbalanced inflow patterns. In the balanced cases, each 4 entry lane had the same mean flow, but for the unbalanced cases, two sets of entry lanes (N-S and E-W) had different combinations of flow patterns without overloading the junction as a whole (see Figure 6.2).

In the following tests, green times for the FT (fixed-time control) were calculated based on Webster's manual methods (2.4). Thus, FT control operates with a set of pre-designed signal plans that use the same stage durations in each cycle. In contrast, VA (the system D Vehicle Actuated control), KCS(D), KCS(D+P), KCS(D+P+TC), Vertical(D), Vertical(D+P) and Vertical(D+P+TC) use varying stage durations between the pre-designed minimum green time and maximum green time.

The resulting performance of each control strategy was compared based on the two sample mean $t-$test and the paired mean $t-$test. For the two chosen control strategies, if the $t-$value of the selected flow level is greater than the critical value $t_\alpha = 1.73$, the difference between two control strategies for that selected flow level is statistically significant at level α =0.05. In this way, if the $t-$value of the paired mean is greater than $t_\alpha = 1.94$, we could conclude with 95% confidence that the second control strategy performs better than the first one over all testing flows . A decision tree for choosing the best control strategy in respect of the paired mean $t-$test results is depicted in Figure 6.8.

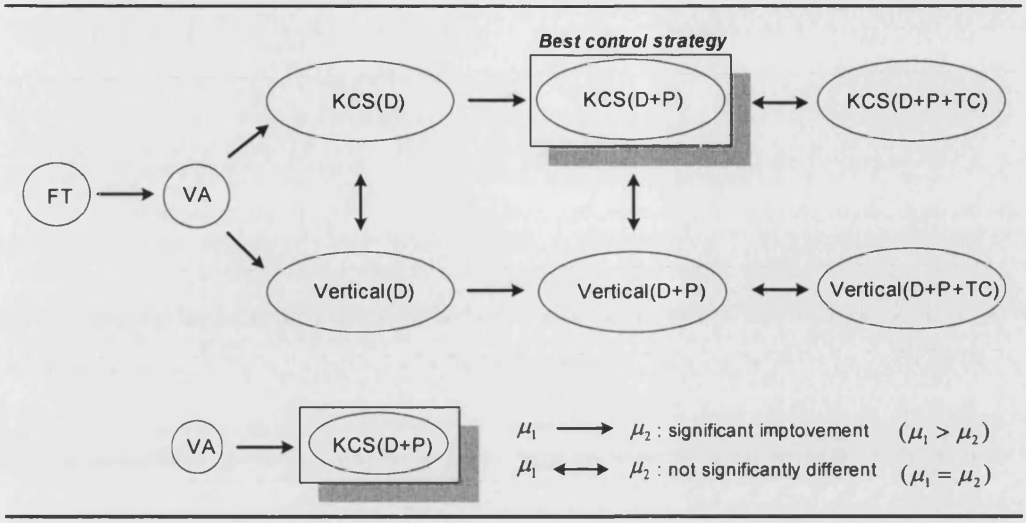


Figure 6.8 A decision tree for choosing the best control strategy according to the paired mean $t-$test results

6.5.1 Results for balanced inflows

In this section, eight different control strategies were tested for seven different conditions of inflows where each entry lane has the same mean flow. Table 6.3 shows the mean rate of delay and associated standard error (S.E.) for 10 runs. This shows that the novel method offers stability in control that is comparable with that of the established methods. The standard errors are in the range between 0.11 and 0.25. We therefore conclude that there is a 95% chance that the sample mean will be within 0.5 vehicles of the population mean. Based on the simulation runs, the results for the two sample mean $t-$test and the paired mean $t-$test are shown in Table 6.4. The paired approach was adopted to eliminate the substantial differences between results that are associated with differences in flow and are common between control methods. The pooled variance for differences between the paired means are in the range between 0.01 and 0.10. At each level of traffic flow, they are not different more than 0.05. Thus, we can see that the range of the pooled variance is similar for each pair of control strategies and they are not substantially different between the methods. The results of the eight different control strategies for the balanced inflows are discussed below:

1) Optimised FT with VA

As would be expected, VA shows better performance than optimised FT. At lower flows the VA is slightly better than FT, but at higher flows VA shows substantially better results. According to the paired mean $t-$test result, we can see that the VA control strategy is better than the optimised FT.

2) VA with KCS(D) and Vertical(D)

Both KCS(D) and Vertical(D) control strategies show the better performance than VA over all flow conditions. For the flow 700 vehicles, both KCS(D) and Vertical(D) control strategies show about 5% less delay than VA. According to the paired mean $t-$test result, we can see that both KCS(D) and Vertical(D) control strategies are better than VA.

3) Vertical(D) with KCS(D)

In most flow conditions, the KCS(D) control strategy is slightly better than the Vertical(D), but overall they are not significantly different. As can be seen in Figure 6.6, these two traffic models show different departure time at the stop-line until the 6^{th} vehicle, which takes about 12.5 seconds from the beginning of the red and amber. When the stage minimum green is 7 seconds and almost certainly at least one vehicle is in the queue before the green starts, the detection period delays estimated from the two control strategies are mostly identical, but the prediction period delay differs because it depends on the number of residual vehicles at the end of each planning period. If they have an intermediate speed and the green comes on whilst the leading vehicle is braking that might affect control decisions. However, such case is unlikely to happen either in test simulations or in practice. From this comparison, we can see that when used for signal timing optimisation, the simpler vertical queueing model is nearly as good as the more detailed KCS one .

4) KCS(D) with KCS(D+P)

Including the prediction period delay, the KCS(D+P) control strategy shows some improvement over most flow conditions. It seems that the prediction period delay does add substantially new information to the objective function. However, it is difficult to define clearly how the prediction period delay is correlated with the detection period delay during the exhaustive searching because the prediction period delay is not proportional to the detection period delay, they fluctuate in each time-step. According to the paired mean $t-$test result, we can see that the KCS(D+P) control strategy is better than the KCS(D).

5) Vertical(D) with Vertical(D+P)

From this test case, it is clear that the prediction period delay is not a big influence for the Vertical queueing model over most flow conditions. However, there is a notable exception for the flow 800 vehicles. In this case, adding the prediction period delay shows a substantial improvement. According to the paired mean $t-$test result, we can see that they are not significantly different.

6) Vertical(D+P) with KCS(D+P)

Including the prediction period delay, the KCS traffic model performs slightly better than Vertical queueing model for most flow conditions. By investigating the simulation results, the prediction period delay estimated from the Vertical queueing model is generally greater than the KCS traffic model due to the presence of the start lag. This is often incurred, for the stream is currently green (see the time t_3 in Figure 5.3.1). According to the paired mean $t-$test result, we can see that KCS(D+P) control strategy is better than Vertical(D+P).

7) KCS(D+P) with KCS(D+P+TC), and Vertical(D+P) with Vertical(D+P+TC)

Terminal cost (TC) does not affect performance at all in these cases. For the flow 800 vehicles, adding TC to the prediction period delay gives slightly more delay for both two control strategies. By investigating the simulation results, the TC value is mostly smaller than the prediction period delay. We could increase the coefficient of the TC function (5.35) to see what difference that makes. However, there will be some correlation between the prediction period delay and the terminal cost; long queues at the end of the planning period will be associated with large values for both of these. In other words, the bigger prediction period delay will be associated with the bigger TC. For this reason, the TC does not add substantially new information to the objective function. According to the paired mean $t-$test result, we can see that they are not significantly different.

8) Best control strategy for the balanced inflow simulation

As can be seen in Tables 6.3 and 6.4, the KCS(D+P) control strategy gives the best performance of all. For the lower flow conditions, it shows 2~3% less delay than VA. In contrast, for the higher flow conditions, it shows about 4~5% less delay than VA. Moreover, Vertical(D+P) performs about as well as KCS(D+P) in these tests.

Table 6.3 Mean rate of delay and associated standard error for balanced flows

Control strategy	Mean rate of delay (vehicles) (standard error)	Flow (vehicles/h)						
		200	300	400	500	600	700	800
FT	Delay (S.E.)	19.91 (0.13)	30.63 (0.16)	41.53 (0.21)	52.77 (0.12)	64.96 (0.13)	77.22 (0.14)	89.46 (0.22)
VA	Delay (S.E.)	19.90 (0.13)	30.56 (0.17)	41.51 (0.17)	52.69 (0.17)	64.43 (0.13)	76.56 (0.16)	89.15 (0.23)
KCS(D)	Delay (S.E.)	19.68 (0.13)	29.96 (0.17)	40.31 (0.15)	50.60 (0.11)	**61.20** (0.13)	**72.63** (0.13)	85.94 (0.25)
KCS(D+P)	Delay (S.E.)	**19.55** (0.13)	**29.92** (0.17)	**40.29** (0.18)	**50.54** (0.12)	**61.20** (0.12)	**72.63** (0.13)	**85.79** (0.25)
KCS(D+P+TC)	Delay (S.E.)	**19.55** (0.13)	**29.92** (0.17)	**40.29** (0.18)	**50.54** (0.12)	**61.20** (0.12)	**72.63** (0.13)	85.94 (0.25)
Vertical(D)	Delay (S.E.)	19.70 (0.13)	29.98 (0.17)	40.33 (0.16)	50.57 (0.11)	**61.20** (0.14)	72.73 (0.11)	86.12 (0.23)
Vertical(D+P)	Delay (S.E.)	19.70 (0.13)	29.97 (0.17)	**40.28** (0.15)	50.60 (0.12)	**61.20** (0.13)	72.73 (0.11)	**85.79** (0.23)
Vertical(D+P+TC)	Delay (S.E.)	19.70 (0.13)	29.97 (0.17)	**40.28** (0.15)	50.60 (0.12)	**61.20** (0.13)	72.73 (0.11)	86.12 (0.23)

Table 6.4 Two sample means t – test and paired mean t – test for balanced flows

Control strategy Comparison			Two sample mean t – test							Paired mean t – test			
μ_1	μ_2		Flow (vehicles/h)							$\bar{D}$	S_D	t – value	Is μ_2 significantly better than μ_1?
			200	300	400	500	600	700	800				
FT	VA	$\mu_1 - \mu_2$	0.01	0.07	0.02	0.08	0.53	0.66	0.31	0.24	0.26	2.40	Yes
		S_p^2	0.02	0.03	0.04	0.02	0.02	0.02	0.05				
		t – value	0.17	0.95	0.23	1.22	9.12	9.82	3.08				
VA	KCS(D)	$\mu_1 - \mu_2$	0.22	0.60	1.20	2.09	3.23	3.39	3.21	2.07	1.44	3.80	Yes
		S_p^2	0.02	0.03	0.03	0.02	0.02	0.02	0.06				
		t – value	3.78	7.89	16.74	32.64	55.56	60.28	29.88				
VA	Vertical(D)	$\mu_1 - \mu_2$	0.20	0.58	1.18	2.12	3.23	3.83	3.03	2.02	1.41	3.81	Yes
		S_p^2	0.02	0.03	0.03	0.02	0.02	0.02	0.05				
		t – value	3.44	7.63	15.98	33.11	53.46	62.38	29.46				
Vertical(D)	KCS(D)	$\mu_1 - \mu_2$	0.02	0.02	0.02	-0.03	0.00	0.10	0.18	0.04	0.07	1.64	No
		S_p^2	0.02	0.03	0.02	0.01	0.02	0.01	0.06				
		t – value	0.34	0.26	0.29	-0.61	0.00	1.86	1.68				
KCS(D)	KCS(D+P)	$\mu_1 - \mu_2$	0.13	0.04	0.02	0.06	0.00	0.00	0.15	0.06	0.06	2.49	Yes
		S_p^2	0.02	0.03	0.03	0.01	0.02	0.02	0.06				
		t – value	2.24	0.53	0.27	1.17	0.00	0.00	1.34				
Vertical(D)	Vertical(D+P)	$\mu_1 - \mu_2$	0.00	0.01	0.05	-0.03	0.00	0.00	0.33	0.05	0.13	1.09	No
		S_p^2	0.02	0.03	0.02	0.01	0.02	0.01	0.05				
		t – value	0.00	0.13	0.72	-0.58	0.00	0.00	3.21				
Vertical(D+P)	KCS(D+P)	$\mu_1 - \mu_2$	0.15	0.05	-0.01	0.06	0.00	0.10	0.00	0.05	0.06	2.23	Yes
		S_p^2	0.02	0.03	0.03	0.01	0.02	0.01	0.06				
		t – value	2.58	0.66	-0.13	1.12	0.00	1.86	0.00				

Table 6.4 Two sample means t – test and paired mean t – test for balanced flows (continued)

Control strategy Comparison			Two sample mean t – test							Paired mean t – test			
μ_1	μ_2		Flow (vehicles/h)							$\bar{D}$	S_D	t – value	Is μ_2 significantly better than μ_1?
			200	300	400	500	600	700	800				
KCS(D+P)	KCS(D+P+TC)	$\mu_1 - \mu_2$	0.00	0.00	0.00	0.00	0.00	0.00	-0.15	-0.02	0.06	-1.00	No
		S_p^2	0.02	0.03	0.03	0.01	0.01	0.02	0.06				
		t – value	0.00	0.00	0.00	0.00	0.00	0.00	-1.34				
Vertical(D+P)	Vertical(D+P+TC)	$\mu_1 - \mu_2$	0.00	0.00	0.00	0.00	0.00	0.00	-0.33	-0.05	0.12	-1.00	No
		S_p^2	0.02	0.03	0.02	0.01	0.02	0.01	0.05				
		t – value	0.00	0.00	0.00	0.00	0.00	0.00	-3.21				
VA	KCS(D+P)	$\mu_1 - \mu_2$	0.35	0.64	1.22	2.15	3.23	3.93	3.36	2.13	1.42	3.95	Yes
		S_p^2	0.02	0.03	0.03	0.02	0.02	0.02	0.06				
		t – value	6.02	8.42	15.58	32.67	57.57	60.28	31.28				
VA	Vertical(D+P)	$\mu_1 - \mu_2$	0.02	0.59	1.23	2.09	3.23	3.83	3.36	2.08	1.44	3.80	Yes
		S_p^2	0.02	0.08	0.07	0.07	0.06	0.06	0.10				
		t – value	3.44	7.76	17.16	31.76	55.56	62.38	32.67				

6.5.2 Results for unbalanced inflows

In this section, eight different control strategies were tested for seven different combinations of unbalanced inflows where two sets of entry lanes (N-S and E-W) have different mean flows without overloading the junction as a whole. Table 6.5 shows the mean rate of delay and associated standard error (S.E.) for 10 runs. The standard errors are in the range between 0.10 and 0.17. We therefore conclude that there is a 95% chance that the sample mean will be within 0.33 vehicles of the population mean. Based on the simulation runs, the results for the two sample mean $t-$test and the paired mean $t-$test are shown in Table 6.6. The pooled variance for differences between the paired means are in the range between 0.01 and 0.03. At each level of traffic flow, they are not different more than 0.02. Thus, we can see that the range of the pooled variance is similar for each pair of control strategies and not substantially different between the methods. The results of the eight different control strategies for the unbalanced inflows are discussed below:

1) Optimised FT with VA

VA shows better performance than optimised FT and the results are slightly better than balanced inflows. According to the paired mean $t-$test result, we can see that the VA control strategy is better than the optimised FT.

2) VA with KCS(D) and Vertical(D)

Both the KCS(D) and the Vertical(D) control strategy show the better performance than the VA over all flow conditions. For the higher flow condition, such as (600, 700), both control strategies show about 5% less delay than VA. According to the paired mean $t-$test result, we can see that both KCS(D) and Vertical(D) control strategies are better than VA.

3) Vertical(D) with KCS(D)

For all flow combinations, there is no significant difference between KCS(D) and Vertical(D) (see discussion in Section 6.4.1 (3)).

4) KCS(D) with KCS(D+P)

Including the prediction period delay, the KCS(D) control strategy shows some improvement over most flow conditions. When each one of lanes has lower flows, such as (200,400) and (200,600), KCS(D+P) performs substantially better than KCS(D). It seems that the prediction period delay is

mainly influential to the lower flow conditions. According to the paired mean $t-$test result, we can see that the KCS(D+P) control strategy is better than the KCS(D).

5) Vertical(D) with Vertical(D+P)

It is clear that that the prediction period delay is not a big influence on the Vertical queueing model over most flow conditions. However, there is some improvement for the flow combinations are (200,400) and (200, 600). According to the paired mean $t-$test result, we can see that they are not significantly different.

6) Vertical(D+P) with KCS(D+P)

Including the prediction period delay, the KCS traffic model performs slightly better than the Vertical queueing model over all flow conditions. According to the paired mean $t-$test result, we can see that the KCS(D+P) control strategy is better than the Vertical(D+P).

7) KCS(D+P) with KCS(D+P+TC), and Vertical(D+P) with Vertical(D+P+TC)

Terminal cost (TC) does not affect these test cases at all (see discussion in Section 6.4.1 (7)).

8) Best control strategy for the unbalanced inflow simulation

As can be seen in Tables 6.5 and 6.6, the KCS(D+P) control strategy gives the best performance of all. In most flow combinations, it shows some improvement over VA. For the flow combination (600,700), the KCS(D+P) control strategy shows about 5% less delay than VA. Moreover, the Vertical (D+P) performs about as well as the KCS(D+P) in these tests.

Table 6.5 Mean rate of delay and associated standard error for unbalanced flows

Control strategy	Mean rate of delay (vehicles) (standard error)	Flow (vehicles/h) (N-S: first row, E-W: second row)						
		200	200	300	300	400	500	600
		400	600	500	600	800	700	700
FT	Delay	30.51	41.13	41.08	46.16	63.70	64.48	71.41
	(S.E.)	(0.17)	(0.13)	(0.16)	(0.13)	(0.12)	(0.14)	(0.13)
VA	Delay	30.47	40.90	40.37	45.64	63.39	64.14	70.60
	(S.E.)	(0.16)	(0.14)	(0.15)	(0.13)	(0.16)	(0.13)	(0.16)
KCS(D)	Delay	29.98	40.51	40.34	45.50	**61.34**	**61.18**	**66.94**
	(S.E.)	(0.16)	(0.15)	(0.15)	(0.12)	(0.10)	(0.15)	(0.15)
KCS(D+P)	Delay	**29.84**	**40.31**	**40.31**	**45.47**	**61.33**	**61.17**	**66.95**
	(S.E.)	(0.17)	(0.16)	(0.18)	(0.15)	(0.10)	(0.14)	(0.12)
KCS(D+P+TC)	Delay	**29.84**	**40.31**	**40.31**	**45.47**	**61.33**	**61.17**	**66.95**
	(S.E.)	(0.17)	(0.16)	(0.18)	(0.15)	(0.10)	(0.14)	(0.12)
Vertical(D)	Delay	29.97	40.48	**40.32**	45.51	61.40	61.21	**66.95**
	(S.E.)	(0.17)	(0.15)	(0.15)	(0.13)	(0.10)	(0.14)	(0.13)
Vertical(D+P)	Delay	29.93	40.34	40.34	45.53	61.42	61.23	66.97
	(S.E.)	(0.16)	(0.16)	(0.17)	(0.14)	(0.12)	(0.14)	(0.14)
Vertical(D+P+TC)	Delay	29.93	40.34	40.34	45.53	61.42	61.23	66.97
	(S.E.)	(0.16)	(0.16)	(0.17)	(0.14)	(0.12)	(0.14)	(0.14)

Table 6.6 Two sample means t – test and paired mean t – test for unbalanced flows

Control strategy Comparison			Two sample mean t – test							Paired mean t – test			
μ_1	μ_2		Flow (vehicles/h) (N-S: first row, E-W: second row)							$\bar{D}$	S_D	t – value	Is μ_2 significantly better than μ_1?
			200	200	300	300	400	500	600				
			400	600	500	600	800	700	700				
FT	VA	$\mu_1 - \mu_2$	0.04	0.23	0.71	0.52	0.31	0.34	0.81	0.42	0.27	4.10	Yes
		S_p^2	0.03	0.02	0.02	0.02	0.02	0.02	0.02				
		t – value	0.54	3.81	10.24	8.94	4.90	5.63	12.42				
VA	KCS(D)	$\mu_1 - \mu_2$	0.49	0.39	0.03	0.14	2.05	2.96	3.66	1.39	1.49	2.47	Yes
		S_p^2	0.03	0.02	0.02	0.02	0.02	0.02	0.02				
		t – value	6.85	6.01	0.45	2.50	34.36	47.16	52.77				
VA	Vertical(D)	$\mu_1 - \mu_2$	0.50	0.42	0.05	0.13	1.99	2.93	3.65	1.38	1.47	2.48	Yes
		S_p^2	0.03	0.02	0.02	0.02	0.02	0.02	0.02				
		t – value	6.77	6.45	0.75	2.24	33.35	48.50	55.99				
Vertical(D)	KCS(D)	$\mu_1 - \mu_2$	-0.01	-0.03	-0.02	0.01	0.06	0.03	0.01	0.01	0.03	0.61	No
		S_p^2	0.03	0.02	0.02	0.02	0.01	0.02	0.02				
		t – value	-0.14	-0.45	-0.30	0.18	1.34	0.46	0.16				
KCS(D)	KCS(D+P)	$\mu_1 - \mu_2$	0.14	0.20	0.03	0.03	0.01	0.01	-0.01	0.06	0.08	1.95	Yes
		S_p^2	0.03	0.02	0.03	0.02	0.01	0.02	0.02				
		t – value	1.90	2.88	0.40	0.49	0.22	0.15	-0.16				
Vertical(D)	Vertical(D+P)	$\mu_1 - \mu_2$	0.04	0.14	-0.02	-0.02	-0.02	-0.02	-0.02	0.01	0.06	0.49	No
		S_p^2	0.03	0.02	0.03	0.02	0.01	0.02	0.02				
		t – value	0.54	2.02	-0.28	-0.33	-0.40	-0.32	-0.33				
Vertical(D+P)	KCS(D+P)	$\mu_1 - \mu_2$	0.09	0.03	0.03	0.06	0.09	0.06	0.02	0.05	0.03	4.99	Yes
		S_p^2	0.03	0.03	0.03	0.02	0.01	0.02	0.02				
		t – value	1.22	0.42	0.38	0.92	1.82	0.96	0.34				

Table 6.6 Two-sample means $t-$test and paired mean $t-$test for unbalanced flows(continued)

Control strategy Comparison			Two sample mean $t-$test							Paired mean $t-$test			
μ_1	μ_2		Flow (vehicles/h) (N-S: first row, E-W: second row)							$\bar{D}$	S_D	$t-$value	Is μ_2 significantly better than μ_1?
			200	200	300	300	400	500	600				
			400	600	500	600	800	700	700				
VA	KCS(D+P)	$\mu_1-\mu_2$	0.63	0.59	0.06	0.17	2.06	2.97	3.65	1.45	1.44	2.65	Yes
		S_p^2	0.01	0.01	0.01	0.01	0.01	0.01	0.01				
		$t-$value	12.45	13.33	1.26	4.14	40.71	72.25	72.14				
VA	Vertical(D+P)	$\mu_1-\mu_2$	0.54	0.56	0.03	0.11	1.97	2.91	3.36	1.39	1.45	2.54	Yes
		S_p^2	0.03	0.02	0.03	0.02	0.02	0.02	0.02				
		$t-$value	7.55	8.33	0.42	1.82	31.15	48.17	53.99				

6.5.3 Sensitivity of control to the maximum green

In this section, sensitivity of control to the choice of maximum green time was tested for the non-optimising method VA and the optimising method KCS(D+P) control strategy. They were compared for the higher flow conditions in the range of 500 to 800 vehicles/h in each lane. A comparison was made for the difference of the mean rate of delay and the average green duration due to the presence of the different maximum green time (e.g. 20 seconds and 30 seconds).

As discussed in Section 2.4.1, the VA control strategy has a tendency to extend greens longer and near its maximum when the maximum green time is set inappropriately. As would be expected, Table 6.7 shows that the VA control strategy gives longer average greens and more delays when the maximum green is set to longer than it needs to be. In contrast, the optimising method by integrating with the KCS traffic model gives reliable solutions, even if the maximum green is set to long (see Figure 6.9 and 6.10). According to these tests, it is clear that this optimising method is not influenced unduly by the pre-defined maximum green time. The results are discussed below:

1) VA with 20 and 30 seconds maximum green

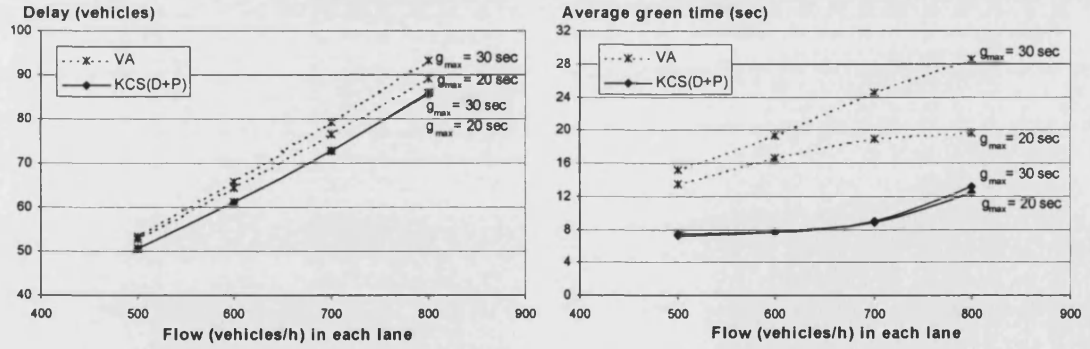
Maximum green time is a big influence on the non-optimising method VA. For the flow condition of 800 vehicles, when the maximum green is set to 30 seconds, there is 4.3% more delay and it takes 9 seconds longer green time than the 20 seconds maximum green. For the flow condition of 500 vehicles, it shows 1.1% more delay and 2 seconds longer green respectively.

2) KCS(D+P) with 20 and 30 seconds maximum green

Maximum green is not a big influence on the optimising method. There is no significant difference between two maximum greens.

Table 6.7 Sensitivity of control to the choice of the maximum green time (mean rate of delay and average green)

Control strategy	Maximum Green (sec)		Flow (vehicles/h)			
			500	600	700	800
VA	20	Delay (vehicles) (S.E.)	52.69 (0.17)	64.43 (0.13)	76.56 (0.16)	89.15 (0.23)
	30	Delay (vehicles) (S.E.)	53.26 (0.17)	65.79 (0.15)	79.31 (0.16)	93.12 (0.31)
	% difference		1.1%	2.1%	3.5%	4.3%
	20 30	Average green (sec)	13.45 15.15	16.60 19.30	18.84 24.51	19.62 28.48
	Average green difference (sec)		1.7	2.7	5.67	8.86
KCS(D+P)	20	Delay (vehicles) (S.E.)	50.54 (0.12)	61.20 (0.12)	72.63 (0.13)	85.79 (0.25)
	30	Delay (vehicles) (S.E.)	50.58 (0.12)	61.19 (0.13)	72.68 (0.13)	86.05 (0.25)
	% difference		0.0%	0.0%	0.0%	0.0%
	20 30	Average green (sec)	7.29 7.27	7.64 7.68	8.83 8.90	12.44 13.18
	Average green difference (sec)		0.0	0.0	0.1	0.7



(a) Mean rate of delay

(b) Average green time

Figure 6.9 Sensitivity of control to the choice of the maximum green time (non-optimising method VA and optimising method KCS)

6.6 DISCUSSION

In this chapter, comparison was made between the various forms of a novel control method with the established System D Vehicle Actuated (VA) and fixed-time (FT) methods. Eight different control strategies were tested for fourteen different combinations of flow patterns (7 balanced and 7 unbalanced flow cases) by using the SIGSIM traffic simulator. Moreover, the performance of six proposed control strategies were compared in terms of the chosen traffic model and the optimiser. The traffic model used here had two choices of trajectory model (KCS traffic model and Vertical queueing model), and three choices of objective function ('Detection period delay', 'Detection period delay + Prediction period delay', or 'Detection period delay + prediction period delay + Terminal Cost'). In contrast, the dynamic optimiser has one rule of operation; a stage-based exhaustive search method in the rolling-horizon concept. The simulation was undertaken for a single isolated junction that has 4 entry links and 2 stages. From these simulation results, it is clear that the optimising method gives better and more reliable performance than the non-optimising method (System D Vehicle Actuated).

As expected, the VA control strategy performs slightly better than FT when all links have equal flows. In addition, VA shows much better improvement when the flows are unbalanced; this is as would be expected because the fixed timings are unbalanced and unresponsive. However, the VA has a tendency to extend green longer and near its maximum when the flow is high. Thus, the VA control strategy gives longer average green and there are more delays when the maximum green is set to longer than it needs to be. In contrast, the optimising method gives reliable solutions even if the maximum green is set for a long period.

A strong improvement can be obtained when the dynamic optimiser uses the KCS traffic model and the objective function uses the detection period and prediction period delays. The first of these corresponds to complete discounting of both prediction period delay and terminal cost, whilst the second to the complete discounting of terminal cost. As noted from the simulation results, it is apparent that the control decision is influenced by the prediction period delay but not by the terminal cost (TC). Thus, we can see that applying the prediction period delay within the lookahead period is effective in the rolling-horizon control concept. Some level of discounting of prediction period delay intermediate between the full inclusion and exclusion might lead to better control, but the results presented here suggest that the scope for improvement in performance is limited. One of the notable finding from these examinations is that the simpler Vertical queueing model is nearly as good as the more detailed KCS traffic model when used for signal timing optimisation.

The present novel optimising method has been interfaced successfully with the SIGSIM traffic simulator. The results shown that this optimising method performs substantially better than the non-optimising VA method. The evaluation results presented in this chapter are SIGSIM estimated performance. If these control strategies are integrated with an other simulation tool, the results might differ, so some further exploration would be desirable.

7. CONCLUSIONS AND SUGGESTIONS FOR FURTHER WORK

7.1 CONCLUSIONS

In this study, a systematic optimising method for use in dynamic signal control was developed. This involved investigating the features of existing control methods by interpreting the detector data, specifying the objective function and making the control decision. This objective was achieved by introducing a novel traffic model and dynamic optimiser. This optimising method was established to have the choice of operational rules in the optimisation framework. The traffic model has two choices of trajectory model (the KCS traffic model and the Vertical queueing model) and three components of distinct delay (detection period delay, prediction period delay and terminal cost). In contrast, the dynamic optimiser has one optimising rule (a stage-based exhaustive search method implemented in the rolling-horizon concept). The key findings of this study responding to the study objectives were summarised as follows:

- Various optimising methods in traffic-responsive signal control have been developed by many authors. The common aim is to take advantage of short-term variations in arrival patterns of the traffic in order to minimise delays, and hence to improve the quality of traffic flow. These approaches are well established, and several of them are in current use. However, there are some features that remain to be investigated: the optimising methods developed in previous studies were subject to various simplifications in respect of one or more rules of operation, in ways of interpreting the detector data, specifying the objective function, and making control decisions.

- Theoretical methods developed for the problem of this optimisation have depended on data concerning arrivals for a considerable time into the future. This simplifies analysis and permits good performance to be achieved in theory, but it is often impractical on the two important grounds of data availability and computational burden. It requires complete information of arrivals over the entire control period and the number of possible combinations of decisions to be calculated is in general too large. On the other hand, practical traffic-responsive methods (non-optimising and optimising methods) use real-time data that are available from vehicle detectors. Non-optimising methods have no specific objective function, but control decisions are made according to certain rules of operation. By contrast, optimising methods use a specific objective function within their optimisation

framework, in which control decisions are made on the basis of the estimates of delay incurred by traffic due to planned signal timings for the lookahead period.

- To overcome the simplicity of the vertical queueing model, a novel traffic model (a Kinematic Car-following model at Signalised junctions: the KCS traffic model) is developed based on the one dimensional kinematic equations of physics. The motion of vehicles from the detector to the stop-line is formulated analytically as a function of the start of green time. It is developed to represent the individual vehicle motion in relation to general car-following concepts, and hence it can apply to the dynamic signal control at microscopic level. The KCS traffic model proposed has some attractions compared to the simpler vertical queueing model. First of all, the KCS traffic model represents the detailed motion of vehicles from the detector to the stop-line, and can provide plenty of information between them. Secondly, the delay calculated from the vertical queueing model is always greater than or equal to the KCS traffic model and started before with respect to variations in the start of green time; however, the maximum sensitivity of delay in the KCS traffic model is greater than that in the vertical queueing model. Thirdly, the vertical queueing model does not consider any braking motion other than an abrupt halt at the stop-line, so that it assumes all vehicles to have the same travel time before joining a queue at the stop-line. Lastly, From these results, the motion of a leading vehicle at the signalised junction is affected by the current signal display, and the motion of following vehicles is dominated by that of the vehicle in front (previous vehicle). Also, it is clear that the leading vehicle trajectory is the most important determinant for estimating the motion of vehicles at signal controlled road junctions.

- A stage-based exhaustive search method in the rolling-horizon concept is proposed and used in the dynamic optimiser. To estimate the operational performance, an objective function is integrated with the traffic model. In this study, the traffic model has two choices of trajectory model (the KCS traffic model and the Vertical queueing model) and three choices of objective function (Detection period delay, Detection period delay + Prediction period delay, or Detection period delay + Prediction period delay + Terminal Cost). At the time of optimisation, based on the pre-defined stage sequences, the exhaustive search method plans a future signal timing (including a duration of lookahead period of about one-cycle) and sends it to the traffic model. The traffic model estimates the delay and sends it back to the dynamic optimiser. When the decision is roll-forward, the current stage is extended by one time-step according to the rolling horizon concept. The key feature in this approach is that an optimal plan is derived with respect to an estimate of delay incurred for the entire lookahead period, but only the first part of the plan is implemented. Hence, the detector data

would be used several times over. This optimising method could identify the best timing plan at the time of optimisation for the objective in hand, though future arrival may render that sub-optimal in the long-run. The lookahead period represents a compromise between the need to consider the future consequences of control decisions implemented in the short term and the limited availability of detailed data concerning vehicular arrivals.

- The novel optimising method is successfully interfaced with SIGSIM to compare the performance difference with two existing control methods: System D Vehicle Actuated (VA) and fixed-time (FT). The main reason for using this kind of traffic simulator is that SIGSIM is an event-based microscopic simulation tool which works on the basis of detectors; any new events can be added and manipulated by the users. Hence, this kind of simulation tool is suitable for a detector based traffic-responsive signal optimisation. One important feature of SIGSIM is that it enables a measure of 'between runs variability', which means that if SIGSIM is set to run more than once with the same input data, including identical random number seed, it will produce identical results. This feature is useful, and it is important when comparing different control strategies that a number of runs are carried out to establish the degree of variation arising from random variations in traffic flow patterns. As part of the final results, SIGSIM provides an operational performance (mean rate of delay) that is qualified by the associated standard error according to the number of runs.

- In the example tests, 8 different control strategies were tested for 14 different combinations of flow patterns (7 balanced and 7 unbalanced) by using the SIGSIM traffic simulator. Each control strategy was run 10 times with the same initial random number seed.

- As expected, the VA control strategy performs slightly better than FT when all links have equal flows. In addition, VA shows much better improvement when the flows are unbalanced; this is as would be expected because the fixed timings are unbalanced and unresponsive. However, the VA has a tendency to extend green longer and near its maximum when the flow is high. Thus, the VA control strategy gives longer average green and there are more delays when the maximum green is set to longer than it needs to be. In contrast, the optimising method gives reliable solutions even if the maximum green is set for a long period.

- A strong improvement can be obtained when the dynamic optimiser uses the KCS traffic model and the objective function uses the detection period and prediction period delays. As

noted from the simulation results, it is apparent that the control decision is influenced by the prediction period delay but not by the terminal cost.

- One of the notable findings from these examinations is that the simpler vertical queueing model is nearly as good as the more detailed KCS traffic model when used for signal timing optimisation.

7.2 SUGGESTIONS FOR FURTHER WORK

The work described so far in this study on the problem of dynamic signal optimisation is only at on early stage, in the sense that there are several potential areas to be further investigated so that more satisfactory results may be obtained. Some suggestions are given as follows:

- For performance evaluation, the optimising method is interfaced with the SIGSIM microscopic traffic simulator as a new control strategy. The results presented in this study are the SIGSIM estimated mean rate of delay. Thus, it is desirable to test the performance of this proposed method with other simulation tools, so some further exploration would be possible.

- The dynamic optimiser developed in this study requires the pre-determined stage sequence and interstage structures, whereas in the phase-based approach only the interstage times need to be specified on the basis of safety considerations. Hence, the phase-based exhaustive search method in the rolling-horizon concept might offer the potential opportunity for a better control performance by proper choice of the stage sequence and the interstage structures. The implementation of the present optimising method in phase-based framework is therefore interesting and could be worthwhile.

- So far, this optimising method is based on minimising the total rate of delay at the junctions. Although this is common in practice, there might exist some situation where optimisation with respect to other performance indices is necessary, such as fuel consumption, emission, or other relevant quantities.

- As noted, the simpler vertical queueing model is nearly as good as the more detailed KCS traffic model when used for the single junction optimisation. However, the performance will

be different if the KCS traffic model is used in network coordinated signal timing optimisation, especially when estimation of the offset is involved.

8. REFERENCES

Allsop, R.E. (1970) Optimisation techniques for reducing delay to traffic in signalised road junction. Ph.D. Thesis, University College London.

Allsop, R.E. (1971) Delay-minimizing settings for fixed-time traffic signals at a single road junction. Journal of the Institute of Mathematics and its Applications, 8 (2), 164-185.

Allsop, R.E. (1972) Estimating the traffic capacity of a signalised road junction. Transportation Research, 6 (3), 245-255.

Allsop, R.E. (1983) Optimisation of timings of traffic signals. Proceedings of the AIRO Conference, Naples, September, 1983, 103-120.

Allsop, R.E. (1992) Evolving application of mathematical optimisation in design and operation of individual signal-controlled road junctions. Proceedings of the IMA Conference on Transport Planning and Control, University of Wales College of Cardiff, September 1989, 57-67. University press, oxford.

Bang, K.L. (1976) Optimal control of isolated traffic signals. Traffic Engineering and Control, 17 (7), 288-292.

Bell, M.G.H., Cowell, M.P.H., and Heydecker, B.G. (1989) Traffic-responsive signal control at isolated junctions. In: Traffic Control methods (eds. S. Yagar and E. Rowe), New York: Engineering Foundation, 273-294.

Bell, M.G.H. (1989) Vehicle responsive traffic signal control. Proceedings of the IMA Conference on Transport Planning and Control, University of Wales, College of Cardiff, September, 253-265..

Bell, M.G.H. and Brookes, D.(1993) Discrete time-adaptive traffic signal control: The calculation of expected delays and stops. Transportation Research C , 1 (1), 43-55.

Breteque, L. and Jazequel, R.(1979) Adaptive control at an isolated intersection- a comparative study of some algorithms. Traffic Engineering and Control, 361-363.

Chandler, R., Herman, R. and Montroll, E. (1958) Traffic dynamics-studies in car following. Operations Research, 6 (2).

Chen, H., Cohen, S.L., Gartner, N.H., and Liu, C.C. (1987) Simulation study of OPAC: A traffic-responsive strategy for traffic signal control. Transportation and Traffic Theory (eds. N.H. Gartner and N.H.M. Wilson), London: Elsevier, 233-249.

Clayton, A.J.H. (1940) Road traffic calculation. J. Inst. Civil Enginerering. 16 (7), 247-284.

COSMOS 2001 (2002) Practical manual for real-time signal control. Seoul Metropolitan Police Agency.

Daganzo, C.F. (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research, 28B (4), 269-287.

Department of Transport (1984) MCE 0141: Microprocessor based traffic signal control for isolated linked and urban traffic control installations. London.

Edie, L.C. (1961) Car following and Steady-state theory for non-congested traffic. Operations Research, 9, 66-76.

FHWA (1982) Model for traffic Operation. Federal highway Administration Report FHWA-TS-82-213, US Department of Transportation.

Gartner, N.H. (1983) OPAC: A demand-responsive strategy for traffic signal control. TRB, Transportation Research Record 906, 75-84.

Gartner, N.H. (2002) Development and implementation of an adaptive control strategy in a traffic signal network: the virtual-fixed-cycle approach. Proceedings of the 15th International Symposium on Transportation and Traffic Theory (eds: Taylor, M.A.P.), 137-155.

Greenberg , H. (1959) An analysis of traffic capacity. Operations research, 7, 78-85.

Gallivan, S. and Heydecker B. (1988) Optimising the control performance of traffic signals at single junction. Transportation Research, 22B (5), 357-370.

Gazis, D.C., Herman, R. and Rothery, R.W. (1961) Nonlinear follow-the-leader models of traffic flow. Operations Research, 9, 545-567.

Gipps, P.G. (1981) A behavioural car-following model for computer simulation. Transportation Research, 15 (2), 105-111.

Greenshields, B.D. (1935) A study of traffic capacity. Highway Research, 14, 448-477.

Henry, J.J., Farges, j.L. and Tuffal, J. (1983) The PRODYN real time traffic algorithm. Proceedings of the 4th IFAC-IFIP-IFORS Conference on Control in transportation Systems, 307-311.

Heydecker, B.G. and Dudgen, I.W. (1987) Calculation of signal settings to minimise delay at a junction. Proceedings of the 10th International Symposium on Transportation and Traffic Theory, MIT, Cambridge, Massachusetts, 159-178.

Heydecker, B.G. (1990) A continuous-time formulation for traffic-responsive signal control. Proceedings of the 11th International Symposium on Transportation and Traffic Theory, Yokohama, 599-618.

Heydecker, B.G., and Boardman R.M. (1999) Optimisation of timings for traffic signals by dynamic programming. UTSG 31st Annual Conference, January 1999, York.

Improta, G. and Cantarella, G.E. (1984) Control system design for an individual signalised junction. Transportation Research B, 18 (2), 147-167.

Institute of Highways and transportation, and Department of transport (1987) Roads and Traffic in Urban Areas. London HMSO.

Kaplan, J.R. and Power, L.D. (1973) Results of SIGOP-TRANSYT comparison studies. Traffic Engineering 43(12) p17-23.

Kimber, R.M., Mcdonald, M., and Hounsell N.B. (1986) The prediction of saturation flows for road junctions controlled by traffic signals. TRRL, Research Report 67.

Law, M. and Crosta, D. (1999) SIGSIM user guide part A: SIGSIM theory. University of Newcastle, Transport operations Research Group and Centre for Transport Studies, University College London.

Lighthill, M.J, and Whitham, G.B. (1955) On kinematic waves II. A theory of flow on long crowded roads. Proceedings of the Royal Society, 229 A, 317-435.

Luenberger, D.G. (1984) Linear and Nonlinear programming. Book, Addison-Wesley.

May, A.D. and Keller, H. (1967) Non-integer car following models. Highway Research Record, 199, 19-32.

Murchland, J.D. (1977) The convexity of Webster's two-term traffic signal delay formula. Note from transport Studies group, University College London (unpublished).

Miller, A. J. (1963) A computer system for traffic networks. Proceedings of the Second International Symposium on the Theory of Road traffic Flow (ed. Almond, J.), OECD, Paris, 200-220.

Nagel, K. and Schreckenberg, M. (1992) A cellular automaton model for freeway traffic. J. Phys. I, 2 2221.

Richard, P.I. (1956) Shockwaves on the highway. Operations Research, 4C (1), 1-12.

Road Research Laboratory (1963) A method of measuring saturation flow at traffic signals. Road Note 34, department of the Environment. Harmondsworth: RRL.

Robertson, D.I. (1967) An improvement to the combination method of reducing delays in traffic networks. RRL Report LR 80.

Robertson, D.I. (1969) TRANSYT: a traffic network study tool. RRL Report, LR253.

Robertson, D.I. and Bretherton, R.D. (1974) Optimum control of an intersection for any known sequence of vehicle arrivals. Proceedings of the 2nd IFAC-IFIP-IFORS Symposium on Traffic Control and Transportation Systems, Monte Carlo.

Robertson, D.I. and Vincent R. A. (1975) Bus propriety in a network of fixed time signals. TRRL Laboratory Report, LR 666.

Sha'Aban (2003) SiGSIM User Guide A, B: Serial SIGSIM User Guide (Version 2.0), University, College London.

Silcock, J.P. (1990) SIGSIN: a phase-based optimisation program for individual signal-controlled junctions. Traffic Engineering and Control, 31 (5), 291-298.

Silcock, J.P. (1993) SIGSIM Version 1.0 Users Guide, Working Paper, University College London, Centre for Transport Studies.

Spiropoulou I. (2003) Modelling blocking back at traffic signals. Ph.D. Thesis, University College London.

Underwood, R. (1961) Speed, volume, and density relationships: Quality and Theory of Traffic Flow. Yale Bureau of Highway Traffic, 141-188.

US Department of Transportation (1991) TRANSYT-7F users guide, FHWA.

Van Zuylean, H. (1976) Acyclic traffic controllers. Note from the transport Studies Group, University College London (unpublished).

Vincent, R.A., Mitchell, A.I., and Robertson, D.I. (1980) User guide to TRANSYT Version 8. TRRL, Laboratory Report 888.

Vincent, R.A. and Young, C.P. (1986) Self-optimising traffic signal control using microprocessor-the TRRL 'MOVA' strategy for isolated intersections. Traffic Engineering and Control, 27 (7/8) 385-7.

Vincent, R.A. and Peirce J.R. (1988) MOVA: Traffic responsive, self-optimising signal control for isolated intersections", TRRL, Research Report 170.

Webster, F.V. (1958): Traffic signal settings. Road Research Technical Paper, 39, London: HMSO.

Webster, F.V. and Cobbe, B.M. (1966) Traffic Signals, Road Research Technical Paper, No.56, London, HMSO.

Whiting, P.D.(1972) The Glasgow traffic control experiment: interim report on SIGOP and TRANSYT. Transport and Road Research Laboratory, LR430.

Wolfram, S. (1986) Theory and applications of Cellular Automata, Singapore: World Science.